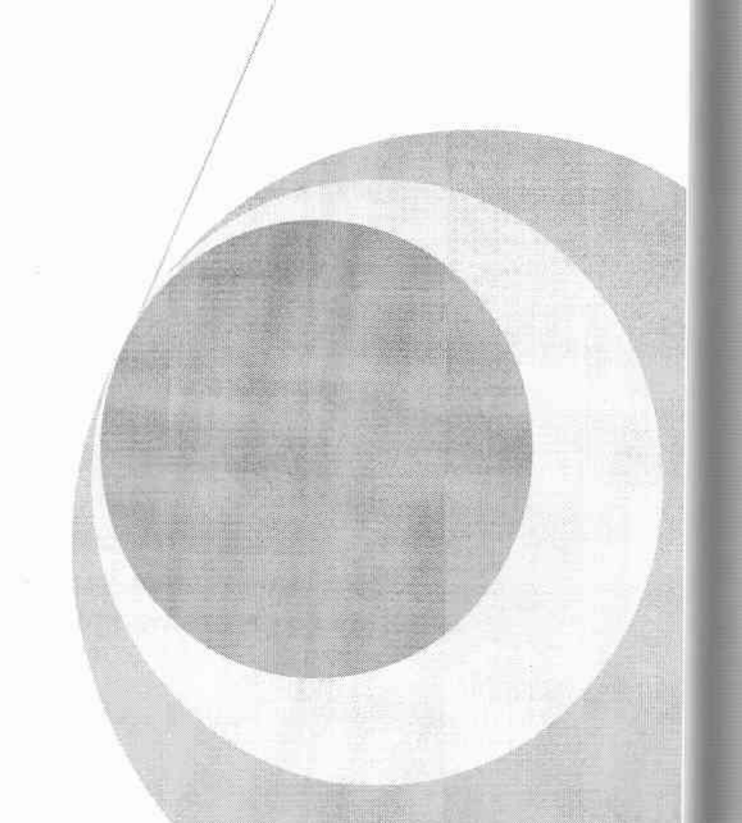


ESTADÍSTICA

Métodos y Aplicaciones

Edwin Galindo



PROCIENCIA EDITORES
2011

Capítulo 1

Análisis Exploratorio de Datos

Nuestra fe en Dios. El resto debe producir datos.
Anónimo

En cualquier actividad de la ciencia, la técnica, los negocios o de la vida cotidiana, que dé como resultado una serie de mediciones, se obtiene más información que las simples cifras recolectadas. El cómo conseguir la información, su análisis e interpretación se puede realizar de muchas maneras, pero primero se debe tener una idea clara de las características más importantes de los datos obtenidos.

Los datos pueden ordenarse en tablas; sin embargo, éstas no muestran su comportamiento global. Su representación gráfica ayuda a captar fácilmente tendencias y establecer modelos probabilísticos. Conjuntamente con el empleo de métodos numéricos sencillos, se puede presentar datos, resumir información y dar una respuesta rápida del comportamiento global de las unidades de donde provienen dichos datos.

En este capítulo examinaremos varios de estos métodos, que son aquellos que frecuentemente aparecen en los paquetes computacionales de estadística.

1.1. Introducción

En primer lugar, demos una definición de la ciencia Estadística que recoge mucho de lo que ella realiza.

La *Estadística* es la ciencia cuyo objetivo es reunir una información cuantitativa concerniente a individuos, grupos, series de hechos, etc. y deducir de ello, gracias al análisis de estos datos, unos significados precisos o unas previsiones para el futuro.

1.1.1. División de la Estadística

Para su mejor estudio, a la Estadística se la divide en dos grandes ramas: la Descriptiva y la Inferencial.

La *Estadística Descriptiva* –también conocida como *Análisis Exploratorio de Datos*– consiste, sobre todo, en la presentación de datos en forma de tablas y gráficos. Está diseñada para resumir o describir los datos sin factores adicionales; esto es, sin intentar inferir nada que vaya más allá de los datos, como tales.

La *Estadística Inferencial* se deriva de muestras, de observaciones hechas sólo acerca de una parte de un conjunto numeroso de elementos y esto implica que su análisis requiere de generalizaciones que van más allá de los datos. Como consecuencia, la característica más importante del reciente crecimiento de la Estadística ha sido un cambio en el énfasis de los métodos que describen a métodos que sirven para hacer generalizaciones. La Estadística Inferencial investiga o analiza una población partiendo de la información obtenida a través de muestras.

1.1.2. Algunos problemas que resuelve la Estadística

Para aplicar los métodos estadísticos a la información disponible, es necesario tener presente los tipos de problemas que esta ciencia resuelve.

Descripción de datos. El primer problema que, históricamente, aborda la Estadística es la descripción de datos. Supongamos que se han tomado ciertas mediciones, que pueden ser los gastos de alimentación en las familias, la producción de las máquinas de un taller, o las preferencias en un grupo de votantes. Se trata de encontrar procedimientos para resumir la información contenida en los datos.

Análisis de muestras. Es frecuente que, por razones técnicas o económicas, no sea posible estudiar los elementos de una población. Por ejemplo, para determinar la opinión de la población ante las elecciones solo se investiga a un grupo pequeño, ya que es imposible consultar a todas las personas en capacidad de votar. Análogamente, se acude a una muestra para estudiar la rentabilidad de un proceso de fabricación o para de terminar el nivel de ocupación de la población.

La Estadística se utiliza para elegir una muestra representativa y para hacer inferencias respecto a la población a partir de lo observado en la muestra. Este es el procedimiento aplicado para, por ejemplo:

- Decidir si un proceso industrial funciona o no adecuadamente, de acuerdo a las especificaciones.
- Estudiar la relación entre consumo de tabaco y cáncer.
- Juzgar respecto a la demanda potencial de un producto, mediante un estudio de mercado.
- Orientar la estrategia electoral de un partido político.
- Interpretar una prueba de inteligencia.

Medición de relaciones. Los gastos en alimentación de una familia dependen de sus ingresos, pero, es imposible determinar con exactitud cuál será el gasto de una familia de ingresos dados. Entonces, no existe una relación exacta, sino estadística. Determinar y medir estas relaciones es importante porque, debido a los errores de medición, las relaciones que observamos entre variables físicas, sociales o técnicas son, casi siempre, estadísticas.

Preguntas como: ¿Depende la calidad de un producto de las condiciones de fabricación y transporte? ¿Cómo se relaciona el rendimiento escolar con variables familiares o sociológicas? ¿Cuál es la relación entre desocupación e inflación?, se responden en términos estadísticos.

Predicción. Muchas variables económicas y físicas tienen cierta inercia en su evolución y aunque sus valores futuros son desconocidos, el estudio de su historia es informativo para prever su comportamiento futuro. Este es el mecanismo que se emplea para prever la demanda de un producto, la temperatura en un horno industrial o las magnitudes macroeconómicas.

1.1.3. Obtención de información

Cuando se examina un proceso o un fenómeno podemos producir una variada información, entonces es preciso determinar cuál es la de interés para los fines que tengamos y cómo conseguirla; así mismo, se debe tener una idea del número de observaciones que son necesarias para disponer de información confiable.

Para la obtención de información estadística se emplean dos formas bien diferenciadas: los métodos de muestreo y los experimentos diseñados.

Una investigación por muestreo es un estudio cuya finalidad es la recolección de datos y en el que el investigador no tiene control sobre las condiciones o los individuos participantes. Ejemplos de muestreos son los censos, las encuestas electorales o de consumo de un producto.

Un experimento es cualquier proceso o estudio en el que se realiza una recolección de datos donde el investigador, usualmente, tiene control sobre algunas de las condiciones bajo las cuales el experimento tiene lugar. Por ejemplo, en el desarrollo de un nuevo medicamento, en la preparación de una nueva aleación de acero para usar en los automóviles, es necesario realizar experimentos para comparar su efectividad con otros previamente existentes.

1.2. Definiciones básicas

Las que antes indicamos son las principales aplicaciones de la Estadística, cuando esta ciencia se utiliza para analizar procesos o fenómenos naturales a profundidad. Pero este no es nuestro caso, por el momento, nosotros podemos pensar que la Estadística es la ciencia de «deducir hechos a partir de datos y de figuras».

Aquí surgen varias ideas importantes en todo análisis estadístico: la *unidad muestral*, la *población* (o *universo*) y la *muestra*.

Definición (de unidad muestral o experimental) Una unidad es una persona, animal, planta o cosa que es examinada por un investigador; es el objeto básico sobre el cual el estudio o experimento se lleva a cabo.

Por ejemplo, una persona, un mono, un plato de semillas, un grupo de facturas.

Definición (de población o universo) Una población es una colección completa de personas, animales, plantas o cosas de las cuales se desea recolectar datos. Es el grupo entero al que queremos describir o del que deseamos sacar conclusiones.

Definición (de muestra) Es un grupo de unidades seleccionadas de la población de acuerdo con un plan o regla, con el objetivo de obtener conclusiones sobre la población de la cual proviene.

El número de unidades que constituyen la muestra se denomina *tamaño muestral.*

Generalmente, se selecciona una muestra porque la población es demasiado grande para estudiarla enteramente. La muestra debe ser representativa de la población general, lo que se logra mediante una selección al azar de las unidades. También, es importante que el investigador defina, completa y cuidadosamente, la población antes de recolectar una muestra, incluyendo una descripción de los miembros a ser seleccionados.

A continuación damos varios ejemplos:

1. Se desea establecer la estructura demográfica, por edad, de la población ecuatoriana. El universo lo forman los datos de nacimientos existentes en las oficinas del Registro Civil. Una muestra puede ser tomada considerando las personas cuyo apellido comienza con la letra A.

2. En un estudio se quiere conocer el «rating» de sintonía de los canales de televisión de una ciudad. La población está constituida por los hogares que poseen televisor y una muestra los hogares de 40 manzanas distribuidas en la ciudad.

3. Una dueña de almacén desea estimar el gasto medio de compra de sus clientes en su almacén en el último año. La población es todas las facturas de compra en el indicado periodo. Una muestra de ciento veinte facturas seleccionadas aleatoriamente, serviría para tener una idea del gasto medio de los clientes.

En los ejemplos anteriores solo se enunciaron posibles muestras para las distintas poblaciones, sin importar que tan buena pudiera ser ésta.[1]

1.3. Datos y escalas de medición

A las mediciones o valores obtenidos en un estudio estadístico se los denomina *datos* provenientes de una variable estadística.

1.3.1. Tipos de datos

Los datos pueden ser:

1. **Cualitativos** (Descriptivos o categóricos): Cuando ellos describen características que no son medibles; por ejemplo, el sexo de un animal, el color de los zapatos, la profesión de una persona.

2. **Cuantitativos** (Numéricos): Cuando ellos describen características que son medibles; por ejemplo, la temperatura del ambiente, el número de hijos de un matrimonio, el salario de una persona.

   A su vez, las variables cuantitativas se clasifican en discretas y en continuas.

   Datos discretos. Un conjunto de datos se denomina discreto si los valores u observaciones que pertenecen a él son distintas y separadas; es decir, ellas pueden ser contadas (1, 2, 3, ...). Ejemplos de datos discretos son: el número de clientes que ingresa a un almacén en un día, el número de años que vive una persona.

   Datos continuos. Un conjunto de datos se denomina continuo si los valores u observaciones que pertenecen a él pueden tomar cualquier valor en un intervalo considerado. Ejemplos de datos continuos son: el tiempo que se demora en ejecutarse un programa en la computadora, el peso de una persona.

1.3.2. Escalas de medición

Definición (de escala de medición) Una escala de medición es un instrumento de medida con el que se asignan valores a las unidades estadísticas.

[1] La elección apropiada de las muestras se explicará en profundidad en el Capítulo 13.

Escala nominal. Un conjunto de datos está medido en *escala nominal* si a los valores que pertenecen a él se les puede asignar un código, en la forma de un número, donde los números son simplemente una etiqueta. Los datos en escala nominal pueden ser contados, pero no pueden ser ordenados o medidos.

Por ejemplo, en un registro de personas, los hombres pueden ser codificados como 0 y las mujeres como 1; el estado civil de un individuo puede codificarse como "1" si es casado y como "2" si no lo es.

Escala ordinal. Un conjunto de datos está medido en *escala ordinal* si a los valores que pertenecen a él se les puede asignar un orden o asociar una escala. Los datos en escala ordinal pueden ser contados y ordenados, pero no pueden ser medidos.

Las categorías, para un conjunto ordinal, deben tener un orden natural; por ejemplo, suponga que a un grupo de personas se les pide que clasifiquen la calidad de la señal de las emisiones de radio, en una escala de 5 a 1, que representan excelente, buena, regular, mala y pésima. Un puntaje de 5 indica mejor señal que un puntaje de 4. Así, los datos resultantes son ordinales.

Escala de intervalo. Un conjunto de datos está medido en *escala de intervalo* si los valores que pertenecen a él pueden tomar cualquier valor dentro de un intervalo finito o infinito, con la particularidad de que existe un «cero relativo». Los datos en escala de intervalo pueden ser contados, ordenados y son válidas las operaciones de adición y sustracción, pero no las de multiplicación y división.

Ejemplos de datos en escala de intervalo son: la temperatura medida en grados centígrados (donde hay un cero elegido arbitrariamente), los puntajes obtenidos en una prueba (donde un puntaje de cero no significa que quien lo obtuvo no sabe nada).

Escala de razón. Un conjunto de datos está medido en *escala de razón* si los valores que pertenecen a él pueden tomar cualquier valor dentro de un intervalo finito o infinito, con la particularidad de que existe un «cero absoluto». Los datos en escala de intervalo pueden ser contados, ordenados y son válidas las operaciones de adición, sustracción, multiplicación y división.

Ejemplos de datos en escala de razón son: la temperatura medida en grados Kelvin (donde hay un cero absoluto), la estatura de una persona, el tiempo de vida útil de una máquina.

1.3.3. Valores atípicos

Un valor atípico –también denominado valor inusual o valor extremo– en un conjunto de datos, es una observación que es lejana, en valor, del resto de datos; es decir, es un dato inusualmente grande o inusualmente pequeño, comparado con los demás.

Un valor atípico puede ser el resultado de un error en una medición, en cuyo caso distorsiona la interpretación de los datos al tener una influencia excesiva sobre los cálculos a partir de la muestra. Si el valor atípico es un resultado genuino es importante, porque podría indicar un comportamiento extremo del proceso en estudio. Por esta razón, todos los valores atípicos deben ser examinados cuidadosamente antes de realizar un análisis formal y no se los debería eliminar sin una justificación previa.

1.4. Características de los datos

Todo conjunto de datos presenta ciertas características que permiten, en una primera aproximación, deducir el comportamiento del proceso del cual fueron obtenidos. Las tres principales características son: la localización, la dispersión y la simetría.

- **Localización.** La localización de un conjunto de datos es la posición relativa que ellos presentan. En general, se mide a la localización por el valor que tiene el punto medio del conjunto de datos.

  Por ejemplo, en la medición de la estatura de un grupo de personas, las mediciones estarán localizadas entre los treinta centímetros (de los recién nacidos) y los dos metros veinte centímetros (de los adultos muy altos), si se supone que estaturas mayores no se presentan, y se puede caracterizar a todos ellos con una estatura promedio de 1.70 metros.

  La idea de localización fue introducida por R. A. Fisher en 1922.

- **Dispersión.** Los valores obtenidos en una muestra no son todos iguales. La variación entre estos valores se denomina *dispersión*. Cuando se mide la dispersión se desea detectar el grado de diseminación de los valores individuales alrededor del centro de las observaciones.

  En los procesos de manufactura o de medición, una alta precisión está asociada con una baja dispersión.

  El concepto de dispersión fue introducido por F. Galton (en 1886) y por W. Lexis (en 1887) e identificado como aquel en el que se reflejan las diferencias entre las mediciones, provenientes de una misma fuente o tomadas en condiciones semejantes.

- **Simetría y asimetría.** Un conjunto de datos es simétrico cuando los valores de los datos están distribuidos en la misma forma por encima y por debajo de su punto medio.

  Los datos simétricos:

  1. Son fáciles de interpretar, pues los datos que están por encima y por debajo del punto medio pueden ser considerados con un mismo criterio;
  2. Permiten la fácil detección de valores atípicos;
  3. Admiten la comparación con conjuntos de datos similares, en términos de la dispersión.

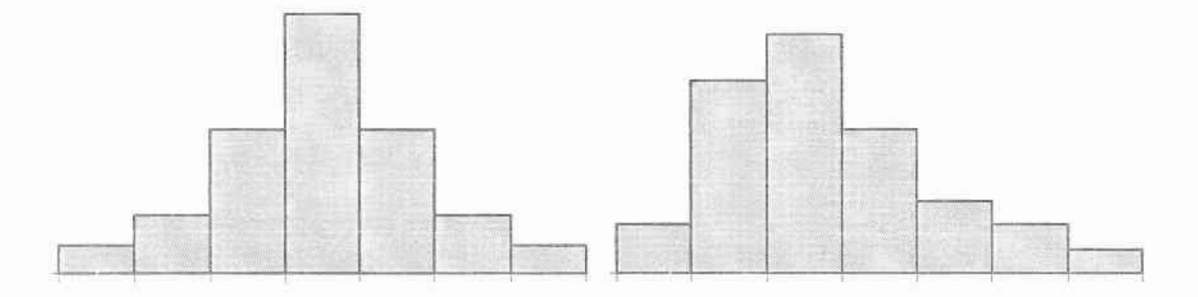


Figura 1.1: Forma esquemática de datos simétricos y asimétricos.

La asimetría en un conjunto de datos es el agrupamiento que ellos presentan a un lado de su centro. Los valores situados a un lado de la mitad de los datos tienden a estar más alejados que los valores que se encuentran en el otro lado.

1.5. Distribución de frecuencias

La *distribución de frecuencias* es una herramienta que se emplea para resumir, mediante una tabla, numerosos datos de manera que se ponga de manifiesto la localización y la dispersión de las observaciones.

Con una tabla de frecuencias se pueden resumir datos categóricos, nominales u ordinales. Si los datos son continuos se puede resumirlos una vez que se los ha dividido en grupos sensibles.

Si se dispone de un número alto de observaciones, n, se procede a establecer cuántas veces se repite cada una de ellas, para determinar su *frecuencia absoluta* n_i. A partir de esta información básica se puede obtener otra, que es conveniente ponerla en una tabla.

Para la confección de una tabla de distribución de frecuencias es recomendable seguir los siguientes pasos:

Procedimiento.

1. Se ordenan los datos $x_1, x_2, \ldots, x_k$ en una columna, de forma ascendente, poniendo a continuación sus frecuencias absolutas $n_1, n_2, \ldots, n_k$. Nótese que $\sum_{i=1}^{k} n_i = n$.

2. Luego se forma una tercera columna en la que se pone la *frecuencia relativa*; que resulta de dividir la frecuencia absoluta n_i para el número total de observaciones: $f_i = \frac{n_i}{n}$. No es más que la proporción de aparecimiento de cada observación.

3. Pueden, también, calcularse dos columnas correspondientes a las frecuencias acumuladas, tanto absoluta como relativa, que resultan de sumar las frecuencias de todas las observaciones anteriores hasta la considerada inclusive. Muchas veces, a las frecuencias relativas se las pone como porcentajes, en lugar de números fraccionarios.

Una tabla de distribución de frecuencias tiene el siguiente aspecto:

Valor de la variable (x_i)	Frecuencia absoluta (n_i)	Frec. absoluta acumulada (N_i)	Frecuencia relativa (f_i)	Frec. relativa acumulada (F_i)
x_1	n_1	$N_1 = n_1$	f_1	$F_1 = f_1$
x_2	n_2	$N_2 = N_1 + n_2$	f_2	$F_2 = F_1 + f_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	$N_k = N_{k-1} + n_k$	f_k	$F_k = F_{k-1} + f_k$
Total		n		1

Ejemplo. En una fábrica de muebles de madera, se controló el tiempo (en minutos) necesario para completar un trabajo de armado de ciertos anaqueles. Se obtuvieron las siguientes mediciones del tiempo empleado por los obreros:

32.9	33.4	33.9	33.2	33.3	32.8	33.1	33.5	33.3	33.5
33.5	33.6	33.7	33.1	33.6	33.4	33.6	33.8	33.9	33.3
33.2	33.6	33.5	33.2	34.4	33.5	33.4	33.1	33.2	33.6
33.5	33.7	32.7	33.8	33.0	33.7	33.1	33.6	33.3	33.5
33.8	33.3	33.4	33.7	34.1	33.2	33.2	33.6	32.9	33.4
32.9	33.9	33.8	33.2	33.5	33.9	34.0	33.5	33.2	33.1
33.1	34.1	33.4	33.6	33.5	33.7	33.7	33.4	33.3	33.4
34.0	33.5	33.0	33.4	33.3	33.4	33.6	33.6	33.7	33.4
33.5	33.6	33.0	33.2	33.1	33.6	33.5	33.6	33.1	33.8
33.7	33.3	33.8	33.1	33.3	33.0	33.3	33.4	33.5	33.0

La siguiente tabla muestra la distribución de frecuencias de datos individuales (en 17 valores).

Tiempo (min)	Frecuencia absoluta (n_i)	Frec. absoluta acumulada (N_i)	Frecuencia relativa (f_i)	Frec. relativa acumulada (F_i)
32.7	1	1	0.01	0.01
32.8	1	2	0.01	0.02
32.9	3	5	0.03	0.05
33.0	5	10	0.05	0.10
33.1	9	19	0.09	0.19
33.2	9	28	0.09	0.28
33.3	10	38	0.10	0.38
33.4	12	50	0.12	0.50
33.5	14	64	0.14	0.64
33.6	13	77	0.13	0.77
33.7	8	85	0.08	0.85
33.8	6	91	0.06	0.91
33.9	4	95	0.04	0.95
34.0	2	97	0.02	0.97
34.1	2	99	0.02	0.99
34.2	0	99	0.00	0.99
34.3	0	99	0.00	0.99
34.4	1	100	0.01	1.00
Total		100		1.00

Se ha presentado una distribución de frecuencias para 100 datos individuales, pero la tabla puede llegar a ser extensa; y si bien presenta la información resumida, puede ser conveniente resumirla aún más, creando clases. La agrupación de datos en clases simplifica la presentación y el estudio de la distribución, aunque se pierden algunos detalles.

A continuación se enumeran los pasos a seguir para construir una distribución de frecuencias de datos agrupados en clases:

1. **Decida el número de clases (k).** La siguiente tabla puede dar una orientación adecuada en la mayoría de los casos.

Número de observaciones	Número de clases recomendado
20 - 50	6
51 - 100	7
101 - 200	8
201 - 500	9
501 - 1000	10
más de 1000	11 - 20

2. **Calcule la longitud de la clase.** La longitud de la clase es igual a la observación mayor menos la menor, dividido por el número de clases. Redondee este resultado para obtener un número conveniente, que tenga el mismo número de decimales que los datos.

$$A = \frac{x_{\text{máx}} - x_{\text{mín}}}{k}$$

3. **Construya las clases indicando los extremos de las mismas.** Como ayuda para cálculos posteriores:

   a) El extremo inferior (L_0) de la primera clase será el número inmediatamente menor al valor mínimo, que tiene un decimal más y que termina en cinco.

b) Los restantes extremos de las clases se obtienen añadiendo repetidamente la longitud de clase al extremo de clase anterior, hasta cubrir todo el rango de valores.

$$L_j = L_{j-1} + A, \quad j = 1, 2, \ldots, k.$$

4. **Marque cada observación dentro de la clase que le corresponda.** Determine la frecuencia absoluta, n_i, correspondiente a cada clase.

5. **Calcule las columnas restantes.** Una vez que tiene la frecuencia absoluta, proceda a calcular las frecuencias relativa y acumuladas, como se explicó anteriormente.

Observación. El número de intervalos puede variar del inicialmente estimado al redondear el valor de la longitud del intervalo y que se cumpla el paso 3 a).

Ejemplo. (Continuación.) Construir una distribución de frecuencias por clases de los datos de las mediciones del tiempo necesario para armar anaqueles.

Solución: De acuerdo a la tabla los datos se distribuirán en $k = 7$ clases. Los máximos y los mínimos son:

$$x_{\text{máx}} = 34.4, \qquad x_{\text{mín}} = 32.7, \qquad x_{\text{máx}} - x_{\text{mín}} = 1.7,$$

$$\text{longitud de la clase } = \frac{1.7}{7} = 0.24,$$

que se redondea a $A = 0.2$.

Fijemos los extremos de los intervalos: el extremo inferior debe ser el número inmediatamente menor al valor mínimo, que termina en 5 y tiene un decimal más que los datos; es decir, $L_0 = 32.65$. Luego, los extremos siguientes se determinan sumando, sucesivamente, 0.2 al extremo inferior hasta sobrepasar el máximo valor de las observaciones:

$$\begin{aligned} L_1 &= L_0 + A = 32.65 + 0.2 = 32.85 \\ L_2 &= L_1 + A = 32.85 + 0.2 = 33.05 \\ &\vdots \\ L_9 &= L_8 + A = 34.25 + 0.2 = 34.45 \end{aligned}$$

Finalmente, se determinan las frecuencias de cada clase.

A continuación se muestran los resultados.

Tiempo (min)	Frecuencia absoluta (n_i)	Frec. absoluta acumulada (N_i)	Frecuencia relativa (f_i)	Frec. relativa acumulada (F_i)
32.65 - 32.85	2	2	0.02	0.02
32.85 - 33.05	8	10	0.08	0.10
33.05 - 33.25	18	28	0.18	0.28
33.25 - 33.45	22	50	0.22	0.50
33.45 - 33.65	27	77	0.27	0.77
33.65 - 33.85	14	91	0.14	0.91
33.85 - 34.05	6	97	0.06	0.97
34.05 - 34.25	2	99	0.02	0.99
34.25 - 34.45	1	100	0.01	1.00
Total		100		1.00

Nótese que por efecto del redondeo en la longitud del intervalo ha dado un total de 9 clases. Queda para el lector realizar el mismo ejercicio redondeando la longitud de la clase a 0.3.

1.6. Representaciones gráficas de los datos

Una manera muy eficiente de conocer el comportamiento de un conjunto de datos es representarlo gráficamente, ya que permite dar una descripción de manera rápida y fácil de entender. La importancia de las representaciones gráficas de los datos es tal, que todo análisis estadístico debe ir acompañado de gráficos, para que el lector pueda entenderlo mejor.

1.6.1. Diagrama de puntos

Un *diagrama de puntos* es una forma de resumir datos cuantitativos, en la que cada observación se representa mediante un punto sobre una recta numérica. Si se dispone de muchos datos, cada punto puede representar un número fijo de individuos.

El diagrama de puntos deja apreciar:

1. La localización general de las observaciones.
2. La dispersión de las observaciones.
3. La presencia de observaciones inusuales o valores atípicos.

Se aconseja utilizar este diagrama para representar hasta un máximo de 20 observaciones individuales, ya que si se dispone de muchos datos es difícil distinguir entre ellos. También, se pueden combinar las representaciones de dos o más conjuntos de datos sobre un mismo gráfico, para lo cual se escogerá una forma de representación que sea de fácil interpretación; por ejemplo, utilizando estrellas (★) o triángulos (▲) en lugar de puntos.

Cuando se construye un diagrama de puntos se deben tomar dos decisiones. La primera es determinar el rango de las observaciones y cómo se lo representará; la segunda es fijar una escala apropiada que permita una buena representación de las observaciones.

Para datos nominales u ordinales, un diagrama de puntos es similar a un gráfico de barras, con las barras reemplazadas por una serie de puntos. Para datos continuos, un diagrama de puntos es similar a un histograma, con los rectángulos reemplazados por puntos. (Véase la sección 1.6.4)

Ejemplo. Las siguientes son las mediciones (en milímetros) de los días de lluvia en el verano de 2006 en el cantón Olmedo de Manabí:

6.4 4.0 3.2 4.6 3.2 8.2 6.0 0.2 4.6 5.2 0.6 2.0 11.8 16.4 3.2.

El diagrama de puntos está dado en la Figura 1.2.

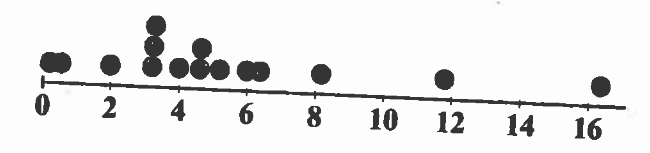


Figura 1.2: Diagrama de puntos.

En el diagrama observamos que:

1. Los datos están agrupados cerca del valor 3, antes que, digamos 8 o 10.
2. Las observaciones se extienden en alrededor de 17 unidades, con una concentración entre 0 y 8.
3. El valor 16.4 puede ser calificado de atípico, porque se encuentra alejado del grupo principal de datos.

1.6.2. Diagrama de tallo y hojas

El diagrama de puntos tiene algunas desventajas: es difícil regresar de los puntos a los datos y puede hacerse confuso si se tiene un número alto de datos. Entonces, es conveniente utilizar otras herramientas para realizar su representación gráfica.

El *diagrama de tallo y hojas*, que es una técnica semigráfica que se emplea para ilustrar las principales características de los datos (localización, dispersión y simetría). Además, tiene la ventaja de presentar los valores de los datos. Por la forma en que se construye, se debe emplear para un conjunto de hasta 100 datos.

Mediante un ejemplo, veamos cómo se realiza el diagrama, paso a paso.

Consideremos los siguientes datos:

08	19	17	01	07	09	05	16
13	04	15	02	00	04	01	12

A los datos los clasificaremos considerando las decenas; así tendremos dos grupos, uno que empieza con 0 y otro que empieza con 1. Ellos forman el tallo, al colocarlos de manera vertical:

0	
1	

A continuación, para cada observación anotamos el segundo dígito (de las unidades) a la derecha de la barra vertical, que vienen a constituir las hojas. La primera observación 08 da

0	8
1	

Al agregar la segunda observación 19, da

0	8
1	9

Y así, se van añadiendo las observaciones hasta obtener:

0	8179542041
1	976352

Los valores que forman las hojas pueden reordenarse de menor a mayor, así:

0	0112445789
1	235679

Podemos crear dos categorías en cada una de las decenas, en las cuales los dígitos de las unidades del 0 al 4 formen un grupo y los dígitos del 5 a 9 formen otro; de esta manera se tiene:

0	142041
0	8795
1	32
1	9765

Cuando los datos constan de más de dos cifras, se deben escoger los rangos para las agrupaciones que se realizarán; luego al llenar las hojas se separan mediante una coma para evitar confusiones. Si disponemos de los siguientes datos:

33	55	79	106	188	47	118	248
47	58	82	113	208	60	88	

Se pueden realizar dos diagramas de tallo y hojas:

0	33, 47, 47, 55, 58, 60, 79, 82, 88
1	06, 13, 18, 88
2	08, 48

que está agrupado por centenas. El siguiente diagrama está agrupado en intervalos de 50:

0	33, 47, 47
0	55, 58, 60, 79, 82, 88
1	06, 13, 18
1	88
2	08, 48
2	

Asimismo, se pueden usar diagramas múltiples para comparar dos conjuntos de datos, para ello se coloca un tallo común y las hojas de un conjunto se ponen a la izquierda del tallo y las hojas del segundo conjunto a la derecha del tallo, de la siguiente manera:

4311	1	44
9888655	1	5779
310	2	0123344
99875	2	678
311	3	03
	3	5
0	4	

Se observa que los datos de la izquierda están más agrupados en los valores bajos, con un rango mayor y fuerte asimetría; mientras que el conjunto de la derecha es muy simétrico y con menor dispersión.

También, se emplean estos diagramas para representar datos con decimales; por ejemplo, si tenemos los datos:

1.3 0.8 1.6 2.0 1.7 1.2 0.5 1.9 0.6 2.2 0.5 1.6.

El diagrama resultante es:

0.	5568
1.	236679
2.	02

1.6.3. Gráfico de sectores y gráfico de barras

Los gráficos de sectores y de barras son dos formas de presentar gráficamente datos categóricos.

Supongamos que los datos aparecen resumidos en una tabla como la siguiente:

Categorías	Frecuencias absolutas (n_i)	Frecuencias relativas (f_i)
C_1	n_1	f_1
C_2	n_2	f_2
$\vdots$	$\vdots$	$\vdots$
C_k	n_k	f_k
Total	n	1

Un *gráfico de sectores* es un círculo dividido en segmentos, donde el área de cada uno de los sectores es proporcional a la frecuencia relativa de esa categoría. El ángulo central de la categoría es igual a $f_i \times 360°$.

Junto a cada uno de los sectores que constituyen el gráfico, se suele indicar el nombre, el número de elementos y el porcentaje de cada categoría.

También, se puede resumir datos cualitativos mediante un *gráfico de barras*. En éstos, los datos se exhiben mediante rectángulos, del mismo ancho, cada uno de los cuales representa una categoría particular. La longitud (y por lo tanto el área) de cada rectángulo es proporcional al número de casos en la categoría que representa.

Si los datos son nominales, las categorías se pueden colocar en cualquier orden; pero si los datos son ordinales, las categorías deben estar ordenadas.

Los gráficos de barras se pueden presentar de manera horizontal o vertical y usualmente hay un espacio entre los rectángulos. Junto a cada uno de los segmentos que componen el gráfico se coloca el nombre, el número de elementos y el porcentaje de cada grupo.

Con el gráfico de barras se distinguen las principales características de los datos, como aquellas causas que son más importantes o que más frecuentemente se presentan en un proceso. También, tiene la ventaja de que se pueden realizar *gráficos de barras agrupadas*, que consiste en representar sobre el mismo gráfico más de dos variables –siempre que estén medidas en las mismas unidades–, permitiendo realizar comparaciones.

Ejemplo. En una empresa financiera, los empleados disponen de computadoras portátiles de distintas marcas. Un resumen del número de máquinas, de acuerdo a su respectiva marca, se presenta en el siguiente cuadro.

Marca	Número de respuestas	%	Marca	Número de respuestas	%
Toshiba	135	42	Lenovo	43	13
Dell	76	23	No sabe	19	6
HP	53	16			

Representar mediante gráficos de sectores y de barras.

Solución: Los gráficos se encuentran en la Figura 1.3.

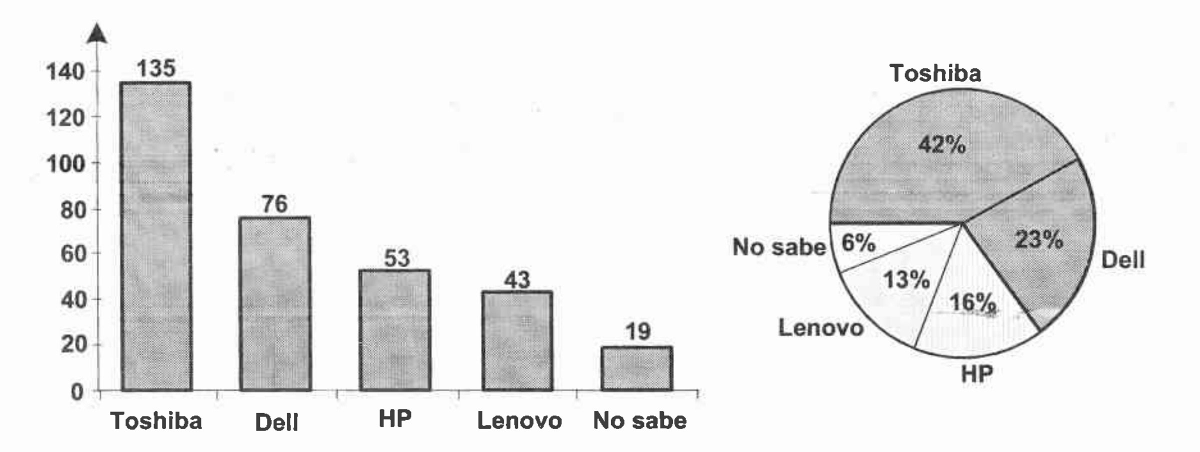


Figura 1.3: Gráficos de barras y de sectores.

1.6.4. Histograma

Un *histograma* es un conjunto de rectángulos, cada uno de los cuales representa un intervalo de agrupación. Sus bases son iguales al intervalo de clase empleado en la distribución de frecuencias y las alturas son proporcionales a la frecuencia absoluta n_i o relativa f_i de la clase.

El histograma es apropiado para datos continuos, medidos con una misma escala y se lo emplea cuando un diagrama de tallo y hojas es tedioso de construir. Igualmente, puede ayudar a detectar observaciones atípicas y cualquier brecha entre los datos.

Ejemplo. (Continuación.) El histograma correspondiente a la tabla de distribución de frecuencias de los tiempos de ensamblaje de anaqueles se presenta a continuación.

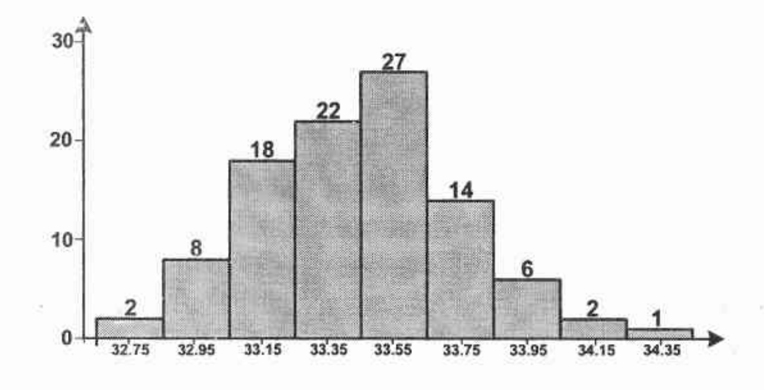


Figura 1.4:

1.6.5. Polígono de frecuencias y ojiva

Un *polígono de frecuencias* es un gráfico que se obtiene uniendo con segmentos de recta los puntos que tienen proporcionalmente como abscisa a la marca de clase y como ordenada la frecuencia respectiva. Se cierra en ambos extremos en las marcas adyacentes con frecuencia cero.

La *ojiva* es un polígono de frecuencias acumuladas; es decir, en las abscisas se colocan los límites superiores de cada intervalo de clase y en las ordenadas se coloca la frecuencia acumulada (absoluta o relativa) de la clase. La ojiva es útil para:

1. Calcular el número o el porcentaje de observaciones que corresponden a un intervalo determinado de la variable.
2. Calcular los percentiles de la distribución de los datos.

Ejemplo. (Continuación.) El polígono de frecuencias y la ojiva, correspondientes a la tabla de distribución de frecuencias de los tiempos de ensamblaje de anaqueles se presenta a continuación.

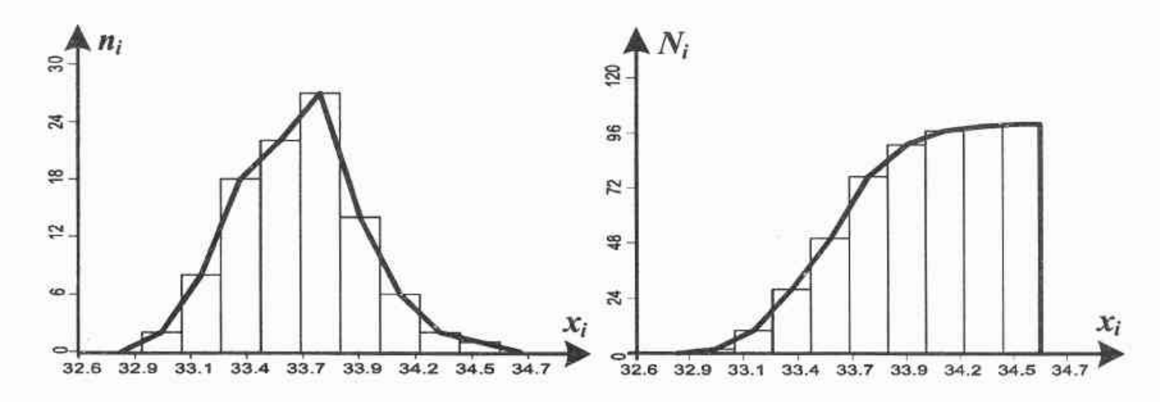


Figura 1.5: Polígono de frecuencias y ojiva.

Una vez que se ha confeccionado una tabla de frecuencias y se ha realizado la representación gráfica correspondiente, es necesario disponer de valores que permitan describir y comparar los conjuntos de datos, mediante números que indiquen su posición, su variabilidad y su forma. Ésto se realiza con las llamadas medidas estadísticas o simplemente estadísticos.

1.7. Ejercicios

1. Dé ejemplos (preferentemente de su propio campo) de poblaciones y muestras.
2. Para cada uno de los distintos tipos de datos: discretos (categóricos, ordinales y nominales) y continuos, enuncie al menos dos ejemplos. Justifique sus respuestas.
3. En una encuesta de opinión acerca de las preferencias de bebidas gaseosas, por sus colores: negro (N), blanco (B) y R (rojo), 20 consumidores dieron las siguientes respuestas:

   N, B, B, N, R, B, B, N, N, B, N, B, B, R, N, B, N, R, N, B.

   Construya el gráfico de sectores circulares.
4. Los siguientes datos corresponden al porcentaje de alumnos de cuarto grado de escuela, clasificados según su rendimiento académico en la materia lenguaje.

Calificación	%
Insuficiente	53
Regular	26
Bueno	15
Muy bueno	5
Sobresaliente	1

a) ¿Con qué tipo de datos está usted trabajando? Explique.

b) Realice los gráficos de pastel y de barras de los datos.

c) ¿Qué porcentaje de los alumnos de cuarto grado tienen un rendimiento «bueno» o mejor que bueno?

5. En la siguiente tabla se describe diferentes razas de perros, según varias características observadas.

Raza	Tamaño	Peso	Velocidad	Agresividad	Función
basset	1	1	1	2	2
boxer	2	2	2	2	1
bauceron	3	2	2	2	3
bulldog	1	1	1	1	1
caniche	1	1	2	1	1
chiguagua	1	1	1	1	1
cocker	2	1	2	2	1
colley	3	2	3	1	1
doberman	3	2	3	2	3
dogo	3	3	3	2	3
fox hound	3	2	3	2	2
galgo	3	2	3	1	2
labrador	2	2	2	1	2
mastin	3	2	3	2	3
pekinés	1	1	1	1	1
podenco	2	2	2	1	2
pointer	3	2	3	1	2
san bernardo	3	3	1	2	3
teckel	1	1	1	1	1
terranova	2	2	1	1	3

donde la codificación es la siguiente:

Tamaño: 1 tamaño pequeño; 2 tamaño mediano; 3 tamaño grande.

Peso: 1 peso pequeño; 2 peso mediano; 3 peso grande.

Velocidad: 1 velocidad leve; 2 velocidad mediana; 3 velocidad grande.

Agresividad: 1 agresividad leve; 2 agresividad grande.

Función: 1 compañía; 2 caza; 3 utilidad.

a) ¿A qué tipo de datos pertenece cada característica definida en la tabla?;

b) Para cada variable, realice el gráfico de pastel o el gráfico de barras;

c) Compare los distintos gráficos y deduzca cuáles variables están relacionadas. Explique su respuesta.

6. Se tiene la siguiente información acerca de la composición del cuerpo humano.

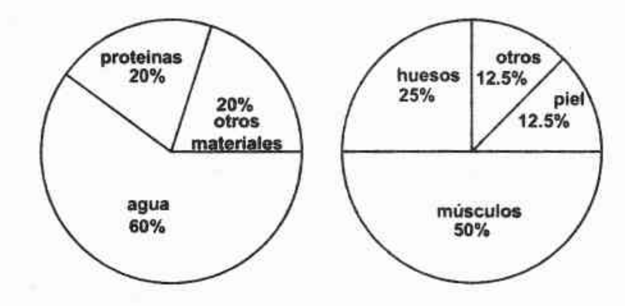


Figura 1.6: Distribución de materiales en el cuerpo y distribución de las proteinas.

¿Qué porcentaje del peso total del cuerpo humano corresponde al peso total de la piel?

7. Se registró la distancia diaria (en km) que el representante comercial de una empresa recorre para visitar a sus clientes:

8.2	13.3	10.1	11.5	13.5	7.6	10.4
4.6	10.5	12.6	13.0	12.0	4.3	7.7
5.9	10.0	10.8	13.1	14.1	5.0	12.0
6.5	12.1	15.0	10.4	13.2	8.3	13.6

a) Realice un diagrama de puntos para los datos;

b) Realice un diagrama de tallo y hojas;

c) Determine la tabla de frecuencias;

d) Dibuje el histograma;

e) Compare este último con los diagramas de puntos y de tallo y hojas.

8. La inversión anual, en miles de dólares, de una muestra de 40 pequeñas empresas fueron:

36	19	29	37	33	22	29	31	21	35
20	4	25	34	24	27	27	24	26	31
27	17	31	10	28	15	41	30	18	39
46	26	12	23	18	33	25	28	23	28

a) Elabore una distribución de frecuencias con 7 intervalos de clase;

b) Realice el diagrama de tallo y hojas;

c) Determine el porcentaje de empresas con una inversión entre 14 mil y 20 mil dólares.

9. Los ingresos mensuales de una muestra de pequeños comerciantes se tabularon en una distribución de frecuencias simétrica de 5 intervalos de clase de igual amplitud, resultando como ingreso mínimo 125 dólares, marca de clase del cuarto intervalo: 300. Si el 8 % de los ingresos son menores que 165 dólares y el 70 % de los ingresos son menores que 275 dólares. ¿Cuál es el porcentaje de los ingresos que son superiores a 285 dólares?

10. Se tiene la siguiente tabla acerca de las edades de los obreros de cierta empresa:

Edades	**No. de obreros**
22 – 27	14
27 – 32	17
32 – 37	25
37 – 42	10
42 – 47	14

Encuentre el porcentaje de obreros cuyas edades están comprendidas entre 35 y 40 años.

11. La siguiente tabla muestra la distribución de las notas en un examen.

Nota	No. alumnos
0 – 5	7
5 – 10	18
10 – 15	15
15 – 20	10

¿Qué porcentaje tuvieron una nota comprendida entre 8 y 17?

12. Al clasificar las notas de 0 a 100 en un examen, se obtuvo una distribución simétrica, con 5 intervalos de clase de igual ancho. Si el 10 % desaprobó con menos de 20, mientras que el 40 % obtuvo notas comprendidas entre 40 y 60, ¿qué porcentaje de alumnos obtuvo una nota menor de 60?

13. En la tabla se indican los tiempos de espera en las ventanillas de un banco.

Tiempo (min)	Frec. absoluta	Frec. relativa
0 – 3	32	
3 – 6		0.30
6 – 9		
9 – 12	8	0.05
12 – 15		0.10

Halle el tamaño de la muestra y complete la tabla de distribución de frecuencias.

14. Los pesos de n artículos se ordenaron en una tabla de distribución de frecuencias de 7 intervalos de igual ancho de clase, donde: mín = 50 g, máx = 120 g.

Además, $f_1 = f_7$, $f_3 = f_5$, $f_5 + f_6 + f_7 = 0.\overline{36}$, $n_1 + n_2 + n_3 + n_4 = 560$ y $n_1 = 64$.

a) Determine el valor de n;

b) ¿Cuántos de estos artículos tendrán un peso mayor o igual a 60 g y menor a 110 g?

15. Halle el tamaño de la muestra y reconstruya la siguiente tabla simétrica de distribución de frecuencias.

Intervalo	Frec. absoluta	Frec. relativa	Frec. relativa acumulada
10 – 12	7		
12 –			0.24
–		0.52	
–	5		
18 – 20			

16. La tabla muestra la distribución del ingreso familiar mensual de 80 familias.

Intervalo	Frec. absoluta	Frec. absoluta acumulada	Frec. relativa
640 – 680			
680 – 720	48	60	
720 – 760			0.125
760 – 800			0.075
800 – 840			

Determine el número de familias que tienen un ingreso menor a 800 dólares mensuales.

17. Dado el siguiente histograma de frecuencias relativas. ¿Cuántas observaciones hay en el rango $[c, f]$, si el total de la muestra es de 400?

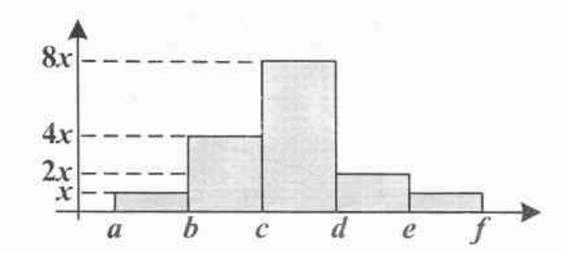


Figura 1.7:

18. En el siguiente gráfico se muestra el consumo de energía en una fábrica.

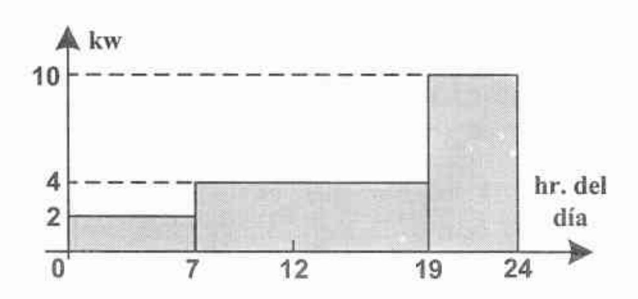


Figura 1.8:

¿Qué porcentaje del consumo diario se utiliza desde las 19h hasta las 24h?

19. En la siguiente ojiva se representan los porcentajes de personas que componen un grupo de personas, según su edad.

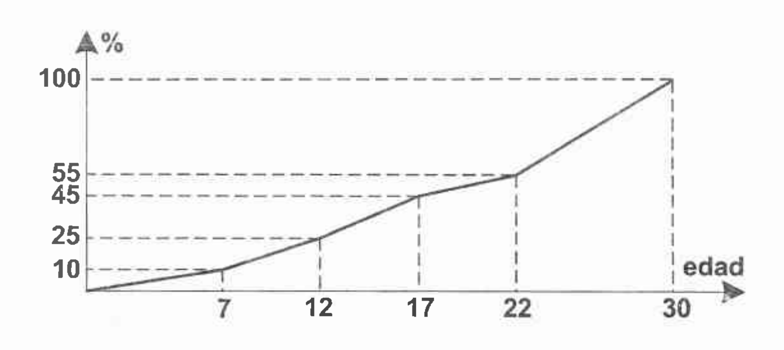


Figura 1.9:

Determine qué porcentaje de personas tienen edades comprendidas entre 10 y 15 años.

20. Dada la ojiva correspondiente a los gastos en servicios de los hogares de una ciudad.

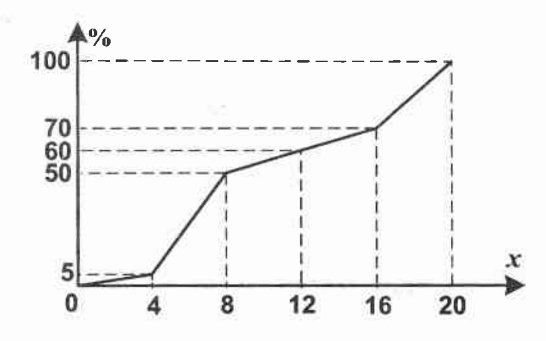


Figura 1.10:

Reconstruya la tabla de distribución de frecuencias.

1.8. Medidas de localización

Cuando se dispone de un conjunto de observaciones, es de interés encontrar el valor en torno al cual se agrupan la mayoría de ellas o el centro de las mismas. Las medidas descriptivas que permiten especificar estos valores se denominan *medidas de localización* o *medidas de tendencia central.*

Existe una amplia variedad de medidas de localización; nos concentraremos en las más empleadas: el promedio, la mediana, la moda, la media geométrica y la media armónica.

1.8.1. La media muestral o promedio

Definición (de promedio o media aritmética) El promedio, notado como $\overline{x}$, de un conjunto de n mediciones $x_1, x_2, \ldots, x_n$ es igual a la suma de sus valores dividido entre n; es decir,

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

- Si las observaciones están agrupadas en una tabla de frecuencias de datos individuales como la siguiente:

Observación	Frec. absoluta
x_1	n_1
x_2	n_2
$\vdots$	$\vdots$
x_k	n_k

donde n_i es la frecuencia absoluta de la observación x_i, el promedio se calcula por

$$\overline{x} = \frac{\sum_{i=1}^{k} n_i x_i}{n}, \quad \text{con} \quad n = \sum_{i=1}^{k} n_i.$$

- Si los datos se presentan en una tabla de frecuencias, agrupados por clases:

Clase	LIC	LSC	Punto medio	Frec. absoluta
1	l_1	s_1	x_1	n_1
2	l_2	s_2	x_2	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	l_k	s_k	x_k	n_k

se calcula el punto medio de cada clase como $x_i = \dfrac{l_i + s_i}{2}$ $(i = 1, 2, \ldots, k)$ y el promedio es

$$\overline{x} = \frac{\sum_{i=1}^{k} n_i\, x_i}{n}, \quad \text{con} \quad n = \sum_{i=1}^{k} n_i.$$

Ventajas e inconvenientes del empleo del promedio:

1. Se expresa en las mismas unidades que la variable.
2. En su cálculo intervienen todos los valores de la distribución.
3. Es el centro de gravedad de toda la distribución, representando a todos los valores observados.
4. Es único.
5. Su principal inconveniente es que se ve afectado por la presencia de valores atípicos.

Ejemplos

1. Calcular el sueldo promedio de diez personas que ganan (en dólares):

170 172 168 165 173 178 180 165 167 172.

Solución: Se dispone de $n = 10$ observaciones sin agrupar, entonces

$$\begin{aligned}\overline{x} &= \frac{x_1 + x_2 + \cdots + x_n}{n} \\ &= \frac{170 + 172 + 168 + 165 + 173 + 178 + 180 + 165 + 167 + 172}{10} \\ &= 171.\end{aligned}$$

2. Calcular la estatura promedio de 46 señoras, cuyas medidas se dan a continuación.

Estatura	1.45	1.48	1.50	1.53	1.55	1.57	1.60	1.63	1.65
Frecuencia	2	4	5	8	12	7	4	3	1

Solución: Como las mediciones están agrupados en una tabla de datos individuales, aplicamos la fórmula que considera la frecuencia de cada una de ellas.

Téngase presente que el número de clases es $k = 9$ y el tamaño de la muestra es $n = 46$.

$$\begin{aligned}\overline{x} &= \frac{\sum_{i=1}^{9} n_i\, x_i}{n} \\ &= \frac{2 \times 1.45 + 4 \times 1.48 + \cdots + 3 \times 1.63 + 1 \times 1.65}{46} \\ &= 1.545.\end{aligned}$$

Los 46 señoras examinadas tienen una estatura promedio de 1.545 metros.

3. En una cooperativa de ahorro y crédito se realizó la tabla de frecuencias de los montos de los ahorros de sus socios (en dólares), según se presenta en la tabla.

Desde	Hasta	Frecuencia
0	100	12
100	200	28
200	30	46
300	400	71
400	500	186
500	600	224
600	700	209
700	800	122
800	900	53
900	1000	19

Calcular el promedio de los ahorros de los socios de la cooperativa.

Solución: Los datos están agrupados en 10 clases. En primer lugar encontraremos el punto medio de cada clase y los pondremos en la tabla:

Desde	Hasta	Punto medio (x_i)	Frecuencia (n_i)
0	100	50	12
100	200	150	28
200	30	250	46
300	400	350	71
400	500	450	186
500	600	550	224
600	700	650	209
700	800	750	122
800	900	850	53
900	1000	950	19

Ahora, empleamos la fórmula que considera la frecuencia de cada una.

Tenemos que $k = 10$ y $\sum_{i=1}^{10} n_i = 970$. Por tanto,

$$\begin{aligned}\overline{x} &= \frac{\sum_{i=1}^{10} n_i\, x_i}{n} \\ &= \frac{12 \times 50 + 28 \times 150 + \cdots + 53 \times 850 + 19 \times 950}{970} \\ &= 555.155.\end{aligned}$$

El ahorro promedio de los cooperados es de 555.16 dólares.

1.8.2. La mediana

La mediana fue por primera vez utilizada, como una medida de localización, por A. A. Cournot en 1843 y redescubierta por F. Galton en 1882, año desde el cual su empleo se ha generalizado.

Definición (de mediana) La mediana de un conjunto de datos $x_1, x_2, \ldots, x_n$ es el valor que se encuentra en el punto medio, cuando se ordenan los valores de menor a mayor.

Se la nota como Q_2 o *Med* y tiene la propiedad de que a cada lado del valor se encuentra el 50 % de las observaciones.

- Si disponemos de un conjunto de datos individuales, para el cálculo de la mediana se procede de la siguiente manera:

  1. Se ordenan las n observaciones $x_1, x_2, \ldots, x_n$ de manera creciente.
  2. Si el número de observaciones es impar, entonces $n = 2m+1$. La mediana es la observación que se encuentra en el lugar $m+1$. Así, si disponemos de $n = 29$ observaciones ordenadas de manera creciente, $m = 14$; es decir, la mediana es la observación que se encuentra en el lugar $14 + 1 = 15$.
  3. Si el número de observaciones es par, entonces $n = 2m$. La mediana es igual a la suma de las observaciones que se encuentran en los lugares m y $m+1$, dividido para dos. Así, si el número de observaciones es de $n = 30$, entonces $m = 15$; la mediana es el promedio de las observaciones que se encuentran en los lugares 15 y 16.

- Si los datos están resumidos en una tabla de distribución de frecuencias de datos individuales.

  1. Ordene las observaciones de manera creciente, con sus respectivas frecuencias acumuladas.
  2. Calcule $\frac{n}{2}$ y redondee al entero más cercano. Determine en la columna de la frecuencia acumulada a qué dato pertenece, comparando el valor obtenido con el valor de la frecuencia acumulada que es igual o inmediatamente superior; éste valor es la mediana.

- Si los datos están resumidos en una tabla de distribución de frecuencias por clases, la mediana se determina por interpolación, así:

  1. Establezca en qué intervalo está el valor mediano. Para ésto, se determina la primera clase cuya frecuencia acumulada sea mayor o igual a $\frac{n}{2}$. Dicho intervalo se denomina *clase mediana.*
  2. La mediana se calcula con la fórmula

  $$Med = L_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i} A,$$

  donde:
  L_{i-1} es el límite inferior de la clase mediana.
  N_{i-1} es la frecuencia acumulada del intervalo inmediatamente anterior al intervalo de la mediana.
  n_i es la frecuencia absoluta de la clase mediana.
  A es la longitud de la clase de la mediana.
  La interpretación gráfica del cálculo de la mediana se encuentra en la Figura 1.11.

Nótese que la mediana de un conjunto de datos no necesariamente pertenece a éste. La propiedad fundamental de la mediana es dividir al conjunto de observaciones en la mitad.

Ventajas e inconvenientes del empleo de la mediana:

1. Es la medida más representativa en el caso de variables que solo admitan la escala ordinal.
2. Es fácil de calcular.
3. En la mediana solo influyen los valores centrales y es insensible a la presencia de valores atípicos.
4. En su determinación no intervienen todos los valores de la variable.

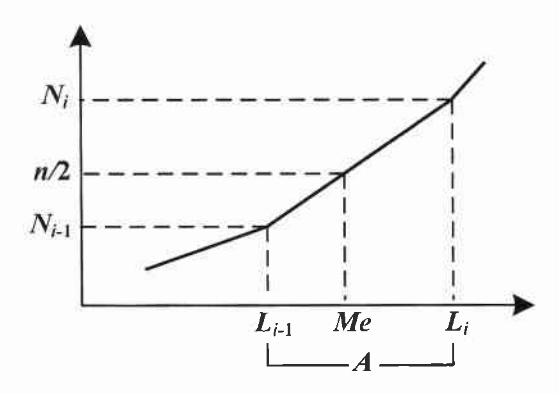


Figura 1.11: Interpretación geométrica del cálculo de la mediana.

Ejemplos

1. Determinar la mediana de los siguientes datos:

   5.5 6.9 7.0 3.0 4.8 4.1 3.2 4.3 5.5 6.5 4.3.

   Solución: Se tienen $n = 11$ observaciones, por lo que $m = 5$, entonces la mediana está en el lugar 5+1. Ordenemos los datos

   3.0 3.2 4.1 4.3 4.3 4.8 5.5 5.5 6.5 6.9 7.0.

   La mediana es la observación que se encuentra en el sexto lugar: $Q_2 = 4.8$.

2. **(Continuación.)** Calcular la mediana de los sueldos de diez personas que ganan (en dólares):

   170 172 168 165 173 178 180 165 167 172.

   Solución: Se tiene $n = 10$ observaciones, que ordenadas dan

   165 165 167 168 170 172 172 173 178 180.

   Por lo tanto, la mediana es el promedio entre las observaciones quinta y sexta:

$$Q_2 = \frac{170 + 172}{2} = 171.$$

3. **(Continuación.)** Calcular la mediana de la estatura de 46 señoras, cuyas medidas son:

Estatura (x_i)	**Frecuencia absoluta** (n_i)	**Frecuencia acumulada** (N_i)
1.45	2	2
1.48	4	6
1.50	5	11
1.53	8	19
1.55	12	31
1.57	7	38
1.60	4	42
1.63	3	45
1.65	1	46

Solución: Las mediciones están agrupados en una tabla de datos individuales y el tamaño de la muestra es $n = 46$.

Calculamos $\frac{n}{2} = 23$ y vemos en la columna de la frecuencia acumulada que hay los valores 19 y 31, que cumplen que $19 \leq 23 < 31$.

Así, la mediana es el valor cuya frecuencia acumulada es 31; es decir, $Q_2 = 1.55\,\text{m}$.

4. Para la liquidación del impuesto a la renta, en una pequeña empresa, se calcularon los ingresos anuales (en dólares) de todos los empleados. La tabla de distribución de frecuencias es la siguiente:

Ingreso anual	**Número de personas (n_i)**	**Frecuencia acumulada (N_i)**
2400 – 3000	3	3
3000 – 4200	20	23
4200 – 5400	35	58
5400 – 7250	25	83
7250 – 9000	15	98
9000 – 12 000	2	100

Solución: Los datos están dados en una tabla de frecuencias por clases con $n = 100$.

Entonces, $\frac{n}{2} = 50$; por tanto, la mediana se encuentra en el intervalo (4200; 5400), de manera que $A = 5400 - 4200 = 1200$.

Ahora, tenemos que

$$\begin{aligned} Med &= L_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i} A \\ &= 4200 + \frac{50 - 23}{35} 1200 = 5125.7. \end{aligned}$$

La mediana del ingreso anual de los empleados de la empresa es 5125.7 dólares.

1.8.3. La moda

Definición (de moda) La moda de un conjunto de datos es aquel valor que tiene la mayor frecuencia absoluta.

Se la nota como *Mo*. Hay ocasiones en las cuales los datos pueden tener dos o más modas, o no puede existir, cuando todos los datos tienen igual frecuencia. Para su determinación es útil construir una tabla de frecuencias de los datos.

- Si los datos están resumidos en una tabla de distribución de frecuencias por clases, la moda se determina mediante la fórmula:

$$Mo = L_{i-1} + \frac{d_1}{d_1 + d_2} A$$

donde:

L_{i-1} es el límite inferior de la clase modal.

d_1 es la diferencia entre la frecuencia de la clase modal y la frecuencia de la clase anterior.

d_2 es la diferencia entre la frecuencia de la clase modal y la frecuencia de la clase siguiente.

A es la longitud de la clase de la mediana.

Aunque la idea de «valor más frecuente» es muy antigua, no fue empleada en estadística, de manera formal, hasta que la popularizó K. Pearson en 1894.

Ventajas e inconvenientes del empleo de la moda:

1. Es fácil de calcular e interpretar.
2. Es la única medida de localización que puede obtenerse en las variables de tipo cualitativo.
3. En su determinación no intervienen todos los valores de la distribución.

Ejemplos

1. Supóngase que las notas de un examen de estadística fueron las siguientes:

   9.4 8.1 9.0 5.6 7.0 9.0 6.5 9.0 3.8 7.0.

   Solución: La moda de este conjunto es Mo=9.0, que es el valor que más veces se repite.

2. Calcular la moda de los siguientes datos:

Observación	2.1	4.5	6.0	8.7	9.2
Frecuencia	5	6	3	2	4

   Solución: La mayor frecuencia es 6, correspondiente al valor 4, por lo tanto Mo=4.

3. Para la liquidación del impuesto a la renta, en una pequeña empresa, se calcularon los ingresos anuales (en dólares) de todos los empleados. La tabla de distribución de frecuencias es la siguiente:

Ingreso anual	**Número de personas** (n_i)
2400 – 3000	3
3000 – 4200	20
4200 – 5400	35
5400 – 7250	25
7250 – 9000	15
9000 – 12 000	2

   Solución: La clase modal es el tercer intervalo, ya que tiene la mayor frecuencia ($n_3 = 35$).

   Entonces, $\frac{n}{2} = 50$; por tanto, la mediana estará en el intervalo (4200; 5400), de manera que $d_1 = 35 - 20 = 15$, $d_2 = 35 - 25 = 10$ y $A = 5400 - 4200 = 1200$.

   Ahora, tenemos que

$$\begin{aligned} Mo &= L_{i-1} + \frac{d_1}{d_1 + d_2} A \\ &= 4200 + \frac{15}{15+10} 1200 = 4920. \end{aligned}$$

   La moda del ingreso anual de los empleados de la empresa es 4920 dólares.

1.8.4. La media geométrica

Definición (de media geométrica) La media geométrica, notada como $\overline{MG}$, de un conjunto de n mediciones $x_1, x_2, \ldots, x_n$ es igual a la raíz $n-$ésima de su producto; es decir,

$$\overline{MG} = \sqrt[n]{x_1 \times x_2 \times \cdots \times x_k}.$$

- Si las observaciones están agrupadas en una tabla de frecuencias de datos individuales,

$$\overline{MG} = \sqrt[n]{x_1^{n_1} \times x_2^{n_2} \times \cdots \times x_k^{n_k}}.$$

- Si las observaciones están agrupadas en una tabla de frecuencias por clases, la expresión es la misma, pero utilizando el punto medio de la clase x_i.

El empleo más frecuente de la media geométrica es el de promediar variables tales como porcentajes, tasas, números índices; es decir, en los casos en los que se supone que la variable presenta variaciones acumulativas.

Ventajas e inconvenientes del empleo de la media geométrica:

1. En su cálculo intervienen todos los valores de la distribución.
2. Los valores extremos tienen menor influencia que en la media aritmética.
3. Es única.
4. Su cálculo es más complicado que el de la media aritmética y solo se la puede calcular cuando todos los valores son positivos.

Ejemplo. Calcular la media geométrica de la estatura de cinco personas que miden (en metros):

1.70 1.72 1.68 1.65 1.73.

Solución: Se dispone de $n = 5$ observaciones; por tanto,

$$\begin{aligned} \overline{MG} &= \sqrt[n]{x_1 \times x_2 \times \cdots \times x_n} \\ &= \sqrt[5]{1.70 \times 1.72 \times 1.68 \times 1.65 \times 1.73} = 1.696. \end{aligned}$$

La media geométrica de las citadas estaturas es 1.696 m.

1.8.5. La media armónica

Definición (de media armónica) La media armónica, notada como $\overline{MH}$, de un conjunto de n mediciones $x_1, x_2, \ldots, x_n$ es el recíproco de la media aritmética de los recíprocos de esos n valores; es decir,

$$\overline{MH} = \frac{n}{\displaystyle\sum_{i=1}^{n} \frac{1}{x_i}}.$$

Su empleo no es aconsejable en distribuciones de variables con valores pequeños. Se suele utilizar para promediar variables tales como productividades, velocidades, tiempos, rendimientos, cambios, etc.

Ventajas e inconvenientes del empleo de la media armónica:

1. En su cálculo intervienen todos los valores de la distribución.
2. Su cálculo no tiene sentido cuando algún valor de la variable toma valor cero.
3. Es única.

Ejemplo. Calcular la media armónica de la estatura de cinco personas que miden (en metros):

$$1.70 \quad 1.72 \quad 1.68 \quad 1.65 \quad 1.73.$$

Solución: Se dispone de $n = 5$ observaciones; por tanto,

$$\begin{aligned}\overline{MH} &= \frac{n}{\displaystyle\sum_{i=1}^{n} \frac{1}{x_i}} \\ &= \frac{5}{\frac{1}{1.70} + \frac{1}{1.72} + \frac{1}{1.68} + \frac{1}{1.65} + \frac{1}{1.73}} = 1.696.\end{aligned}$$

La media armónica de las citadas estaturas es 1.696 m.

1.8.6. Percentiles, cuartiles y quintiles

Antes de finalizar esta sección, es conveniente referirnos a varios términos que son de uso común en la práctica estadística: los cuartiles, los quintiles y los percentiles. Estas medidas estadísticas corresponden a lo que se denomina *medidas de posición no central.*

A un conjunto de datos ordenado se lo puede dividir en un número fijo de partes iguales; cuando se lo divide en cien partes se tienen los percentiles.

Definición (de percentiles) Los percentiles son cada uno de los 99 valores que dividen a la distribución de los datos en 100 partes iguales.

A los percentiles se les nota como P_k. Con ellos se puede encontrar regiones donde se acumulan los datos; así, el 30 % de los datos están por debajo del trigésimo percentil.

Para su cálculo se procede de la siguiente manera:

- Si los datos no están agrupados o están en una tabla de datos individuales, se efectúa la siguiente descomposición:

$$\frac{nk}{100} = j + r,$$

donde:

j es la parte entera de $\frac{nk}{100}$.

r es la parte fraccionaria de $\frac{nk}{100}$.

Entonces, se tiene que

$$P_k = \begin{cases} \frac{x_j + x_{j+1}}{2}, & \text{si } r = 0; \\ x_{j+1}, & \text{si } r > 0. \end{cases}$$

- Si los datos están agrupados en clases, se calcula mediante

$$P_k = L_{k-1} + \frac{\frac{nk}{100} - N_{k-1}}{n_k} A,$$

donde:

L_{k-1} es el límite inferior del intervalo h (cuya frecuencia acumulada es la primera mayor o igual a $\frac{nk}{100}$).

N_{k-1} es la frecuencia acumulada hasta L_{k-1}.

n_k es la frecuencia absoluta del intervalo h.

A es la longitud del intervalo h.

Ejemplos

1. Calcular los percentiles de orden 20 y 33 de la estatura de diez personas que miden (en cm):

165 165 167 168 170 172 172 173 178 180.

Solución: Tenemos que $n = 10$.

- Para P_{20}, $k = 20$

$$\begin{aligned} \frac{nk}{100} &= j + r \\ \frac{10 \times 20}{100} &= 2 + 0. \end{aligned}$$

Entonces, $r = 0$ y $j = 2$:

$$\begin{aligned} P_k &= \frac{x_j + x_{j+1}}{2} \\ P_{20} &= \frac{165 + 167}{2} = 166. \end{aligned}$$

- Para P_{33}, $k = 33$

$$\frac{10 \times 33}{100} = 3 + 0.3.$$

Entonces, $r = 0.3$ y $j = 3$.

$$\begin{aligned} P_k &= x_{j+1} \\ P_{33} &= x_4 = 168. \end{aligned}$$

2. **(Continuación.)** Calcular el percentil de orden 86 de los ingresos anuales de los empleados de una empresa.

Ingreso anual	**Número de personas (n_i)**	**Frecuencia acumulada (N_i)**
2400 – 3000	3	3
3000 – 4200	20	23
4200 – 5400	35	58
5400 – 7250	25	83
7250 – 9000	15	98
9000 – 12 000	2	100

Solución: Tenemos que $n = 100$.

- Para P_{86}, $k = 86$ y $\dfrac{nk}{100} = \dfrac{100 \times 86}{100} = 86$.
  El intervalo h donde se encuentra P_{86} es $(7250, 9000)$ y $L_{k-1} = 7250$.
  También, se tiene que $N_{k-1} = 83$, $n_k = 15$ y $A = 9000 - 7250 = 1750$.
  Con estos datos, obtenemos:

$$
\begin{aligned}
P_k &= L_{k-1} + \frac{\frac{nk}{100} - N_{k-1}}{n_k} A \\
P_{86} &= 7250 + \frac{86 - 83}{15} 1750 \\
&= 7600. \quad \blacktriangleleft
\end{aligned}
$$

Dos casos particulares, y muy utilizados, resultan cuando al conjunto de datos se lo divide en cuatro o cinco partes iguales, que corresponden a los cuartiles y a los quintiles, respectivamente.

Definición (de cuartiles) Son valores que dividen a la distribución de los datos en 4 partes, cada una de las cuales engloba el 25 % de los mismos.

Los cuartiles son 3:

- El cuartil inferior (Q_1), que deja a su izquierda el 25 % de los datos y se cumple que $Q_1 = P_{25}$.
- El cuartil medio (Q_2), que deja a su izquierda el 50 % de los datos, coincide con la mediana y se cumple que $Q_2 = P_{50}$.
- El cuartil superior (Q_3), que deja a su izquierda el 75 % de los datos y se cumple que $Q_3 = P_{75}$.

Así, para el cálculo de los cuartiles solo se deberá tener en cuenta que ellos son los percentiles de orden 25, 50 y 75, respectivamente (Figura 1.12).

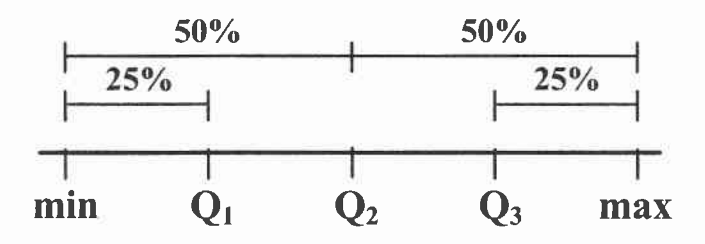


Figura 1.12: Disposición de los cuartiles en un conjunto de datos.

Definición (de quintiles) Los quintiles son valores que dividen a la distribución de los datos en cinco grupos, cada uno de los cuales contiene el 20 % de las observaciones.

Los quintiles son 4:

- El primer quintil (q_1), que deja a su izquierda el 20 % de los datos y se cumple que $q_1 = P_{20}$.
- El segundo quintil (q_2), que deja a su izquierda el 40 % de los datos y se cumple que $q_2 = P_{40}$.
- El tercer quintil (q_3), que deja a su izquierda el 60 % de los datos y se cumple que $q_3 = P_{60}$.
- El cuarto quintil (q_4), que deja a su izquierda el 80 % de los datos y se cumple que $q_4 = P_{80}$.

Ejemplos

1. **(Continuación.)** Determinar los cuartiles inferior y superior de las estaturas de 46 señoras, cuyas medidas son:

Estatura (x_i)	**Frecuencia absoluta** (n_i)	**Frecuencia acumulada** (N_i)
1.45	2	2
1.48	4	6
1.50	5	11
1.53	8	19
1.55	12	31
1.57	7	38
1.60	4	42
1.63	3	45
1.65	1	46

Solución: Tenemos que $n = 46$.

- Para el cuartil inferior, $Q_1 = P_{25}$, por tanto, $k = 25$ y

$$\begin{aligned} \frac{nk}{100} &= t + r \\ \frac{46 \times 25}{100} &= 11 + 0.5. \end{aligned}$$

De manera que, $r = 0.5$ y

$$\begin{aligned} P_k &= x_{j+1} \\ P_{25} &= x_{12} = 1.53. \end{aligned}$$

- Para el cuartil superior, $Q_3 = P_{75}$, $k = 75$ y

$$\begin{aligned} \frac{nk}{100} &= t + r \\ \frac{46 \times 75}{100} &= 35 + 0.5. \end{aligned}$$

Es decir, $r = 0.5$ y

$$\begin{aligned} P_k &= x_{j+1} \\ P_{75} &= x_{36} = 1.57. \end{aligned}$$

2. **(Continuación.)** Determinar los cuartiles inferior y superior de los ingresos anuales de los empleados de una empresa.

Ingreso anual	**Número de personas** (n_i)	**Frecuencia acumulada** (N_i)
2400 – 3000	3	3
3000 – 4200	20	23
4200 – 5400	35	58
5400 – 7250	25	83
7250 – 9000	15	98
9000 – 12 000	2	100

Solución: Tenemos que $n = 100$.

- **Cuartil inferior**: $Q_1 = P_{25}$, $k = 25$ y $\dfrac{nk}{100} = \dfrac{100 \times 25}{100} = 25$.
  El intervalo h donde se encuentra Q_1 es (4200; 5400) y $L_{k-1} = 4200$.
  También, se tiene que $N_{k-1} = 23$, $n_k = 35$ y $A = 5400 - 4200 = 1200$.
  Entonces, resulta que:

$$\begin{aligned} P_k &= L_{k-1} + \frac{\frac{nk}{100} - N_{k-1}}{n_k} A \\ P_{25} &= 4200 + \frac{25-23}{35} 1200 \\ &= 4268.6. \end{aligned}$$

- **Cuartil superior**: $Q_3 = P_{75}$, $k = 75$ y $\dfrac{nk}{100} = \dfrac{100 \times 75}{100} = 75$.
  El intervalo h donde se encuentra P_{75} es (5400; 7250) y $L_{k-1} = 5400$.
  También, se tiene que $N_{k-1} = 58$, $n_k = 25$ y $A = 7250 - 5400 = 1850$.
  Con estos datos, obtenemos:

$$\begin{aligned} P_k &= L_{k-1} + \frac{\frac{nk}{100} - N_{k-1}}{n_k} A \\ P_{75} &= 5400 + \frac{75-58}{25} 1850 \\ &= 6658. \end{aligned}$$

1.9. Medidas de dispersión

Luego de determinar la localización de las observaciones, es conveniente medir su grado de dispersión alrededor del centro. Las medidas que permiten especificar esta característica se denominan *medidas de dispersión*.

Estas medidas deben tener la propiedad de que si los datos están ampliamente extendidos, la medida será alta; y cuando los datos se encuentren muy agrupados, será baja.

Existen varias medidas de dispersión, nosotros vamos a analizar la desviación estándar, el rango y el rango intercuartil.

1.9.1. La desviación estándar

Una vez que se ha calculado el promedio de las mediciones, un indicador de su variabilidad es la desviación de cada medición particular con respecto al promedio, $x_i - \overline{x}$. Pero ésta da una información válida para cada medición y no para toda la muestra. Para tal efecto se emplea la desviación estándar, medida de dispersión fue introducida por K. Pearson en 1894.

Definición (de desviación estándar o desviación típica) La desviación estándar, notada como s, de un conjunto de n mediciones x_1, x_2, ..., x_n es la raíz cuadrada de la suma de los cuadrados de las desviaciones de las mediciones, respecto al promedio $\overline{x}$, dividida entre $n - 1$; es decir,

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2}.$$

Nótese que la desviación estándar es siempre positiva y sus unidades de medida son las mismas que aquellas que corresponden a los datos originales.

Para su cálculo también se emplea la fórmula equivalente

$$s = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\,(\overline{x})^2}{n-1}}.$$

De la misma manera que para la media aritmética se consideran los siguientes casos:

- Si las observaciones están agrupadas en una tabla de frecuencias de datos individuales:

Observación	Frec. absoluta
x_1	n_1
x_2	n_2
$\vdots$	$\vdots$
x_k	n_k

la desviación estándar se calcula por

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{k} n_i(x_i - \overline{x})^2} \quad \text{o} \quad s = \sqrt{\frac{\sum_{i=1}^{k} n_i\, x_i^2 - n\,(\overline{x})^2}{n-1}}, \quad \text{con} \quad n = \sum_{i=1}^{k} n_i.$$

- Si los datos se presentan en una tabla de frecuencias, agrupados por clases:

Clase	LIC	LSC	Punto medio	Frec. absoluta
1	l_1	s_1	x_1	n_1
2	l_2	s_2	x_2	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	l_k	s_k	x_k	n_k

s se calcula por

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{k} n_i(x_i - \overline{x})^2} \quad \text{o} \quad s = \sqrt{\frac{\sum_{i=1}^{k} n_i\, x_i^2 - n\,(\overline{x})^2}{n-1}}, \quad \text{con} \quad n = \sum_{i=1}^{k} n_i.$$

Ventajas e inconvenientes del empleo de la desviación estándar:

1. Se expresa en las mismas unidades que los datos originales.
2. En su cálculo intervienen todos los valores de la distribución y por ello puede ser complicado.
3. Es única.
4. Se ve muy afectada por la presencia de valores atípicos.

Ejemplos

1. **(Continuación.)** Calcular la desviación estándar de los sueldos de diez personas que ganan (en dólares):

   170 172 168 165 173 178 180 165 167 172.

   Solución: Previamente se había calculado el promedio $\overline{x} = 171$. Con ésto, resulta que:

$$\begin{aligned} s &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \\ &= \sqrt{\frac{(170-171)^2 + (172-171)^2 + \cdots + (167-171)^2 + (172-171)^2}{10-1}} \\ &= 5.1. \end{aligned}$$

   Esos sueldos tienen una desviación estándar de 5.1 dólares.

2. **(Continuación.)** Calcular la desviación estándar de la estatura de 46 señoras, cuyas medidas son:

Estatura	1.45	1.48	1.50	1.53	1.55	1.57	1.60	1.63	1.65
Frecuencia	2	4	5	8	12	7	4	3	1

   Solución: Anteriormente se determinó que $n = 46$, $k = 9$ y $\overline{x} = 1.545$.

   Para realizar el cálculo, obtengamos el valor de $\sum_{i=1}^{k} n_i\, x_i^2$:

$$\sum_{i=1}^{k} n_i\, x_i^2 = 2(1.45)^2 + 4(1.458)^2 + \cdots + 3(1.63)^2 + 1(1.65)^2 = 109.9615.$$

   Entonces, se tiene que

$$s = \sqrt{\frac{\sum_{i=1}^{k} n_i\, x_i^2 - n\,(\overline{x})^2}{n-1}} = \sqrt{\frac{109.9615 - 46(1.545)^2}{46-1}} = 0.04627.$$

   La estatura de las señoras analizadas tiene una desviación estándar de 4.6 cm.

3. **(Continuación.)** Calcular la desviación típica de los montos de ahorros de los socios de una cooperativa de ahorro y crédito:

Desde	**Hasta**	**Punto medio (x_i)**	**Frecuencia (n_i)**
0	100	50	12
100	200	150	28
200	30	250	46
300	400	350	71
400	500	450	186
500	600	550	224
600	700	650	209
700	800	750	122
800	900	850	53
900	1000	950	19

   Solución: **Antes se determinó que $n = 970$, $k = 10$ y $\overline{x} = 555.155$.**

Calculemos la siguiente sumatoria:

$$\sum_{i=1}^{9} n_i x_i^2 = 12(50)^2 + 28(150)^2 + \cdots + 53(850)^2 + 19(950)^2 = 330\,025\,000.$$

De manera que la desviación típica es

$$s = \sqrt{\frac{\sum_{i=1}^{k} n_i x_i^2 - n(\overline{x})^2}{n-1}} = \sqrt{\frac{330\,025\,000 - 970(555.155)^2}{970-1}} = 179.08. \quad \blacktriangleleft$$

Conjuntamente con la desviación estándar se suele definir la *varianza muestral* de un conjunto de datos, notada s^2, como la suma de los cuadrados de las desviaciones respecto a su promedio, dividido por el uno menos que el número de observaciones en el conjunto de datos y se calcula mediante

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

1.9.2. El rango y el rango intercuartil

Definición (de rango o recorrido) El rango de n mediciones es igual a la diferencia entre los valores mayor y menor de las mismas:

$$\text{Rango} = x_{\text{máx}} - x_{\text{mín}}.$$

El rango se puede utilizar para hallar una aproximación de la desviación estándar mediante las siguientes relaciones:

$$s \approx \frac{\text{Rango}}{\sqrt{n}}, \quad \text{para } n \leq 16, \qquad s \approx \frac{\text{Rango}}{4}, \quad \text{para } 16 < n \leq 100,$$

$$s \approx \frac{\text{Rango}}{5}, \quad \text{para } 100 < n \leq 400, \qquad s \approx \frac{\text{Rango}}{6}, \quad \text{para } n > 400.$$

Ventajas e inconvenientes del empleo del rango:

1. En su cálculo solo intervienen los dos valores extremos de la distribución y por ello se ve muy afectado por la presencia de valores atípicos.

2. Es fácil de calcular.

Definición (de rango intercuartil) El rango intercuartil, notado por RIQ, de un conjunto de datos es igual a la diferencia entre los cuartiles superior e inferior; es decir,

$$RIQ = Q_3 - Q_1.$$

Las definiciones de los cuartiles superior e inferior y del rango intercuartil fueron dadas por F. Galton en 1882.

Ventajas e inconvenientes del empleo del rango intercuartil:

1. Es fácil de calcular.
2. Se ve poco afectado por la presencia de valores atípicos.
3. En su determinación no interviene la totalidad de los datos.

Ejemplo. (Continuación.) Calcular la desviación estándar de la estatura de 46 señoras, cuyas medidas se resumen en la siguiente tabla:

Estatura	1.45	1.48	1.50	1.53	1.55	1.57	1.60	1.63	1.65
Frecuencia	2	4	5	8	12	7	4	3	1

Solución: Antes se determinó que $Q_1 = 1.53$ y $Q_3 = 1.57$. Además, $x_{\text{mín}} = 1.45$ y $x_{\text{máx}} = 1.65$.

Entonces,

$$\begin{aligned} \text{Rango} &= x_{\text{máx}} - x_{\text{mín}} = 1.65 - 1.45 = 0.20. \\ RIQ &= Q_3 - Q_1 = 1.57 - 1.53 = 0.04. \end{aligned}$$

Además, podemos calcular una aproximación de la desviación estándar de los datos:

$$s \approx \frac{\text{Rango}}{4} = \frac{0.20}{4} = 0.05.$$

Como se ve, el valor aproximado es bastante cercano al exacto, calculado con la fórmula respectiva.

1.9.3. El coeficiente de variación

Definición (de coeficiente de variación) El coeficiente de variación, notado por CV, es igual a la desviación estándar dividida por la media aritmética; es decir,

$$CV = \frac{s}{\bar{x}}.$$

Esta medida se utiliza para comparar las mediciones de una misma magnitud realizada en distintas unidades o por distintos individuos.

Si un conjunto de datos es homogéneo, $CV \leq 1$; si $CV > 1.5$, los datos podrían ser heterogéneos.

Ventajas e inconvenientes del empleo del coeficiente de variación:

1. Es una medida adimensional.
2. En su cálculo intervienen todas las observaciones, pudiendo ser muy influido por valores atípicos.
3. Puede ser difícil de interpretar.
4. Pierde su significado si el promedio es igual a cero.

Ejemplo. (Continuación.) Calcular el coeficiente de variación del sueldo de diez personas que ganan (en dólares):

170 172 168 165 173 178 180 165 167 172.

Solución: Previamente se había calculado que $\overline{x} = 171$ y $s = 5.1$. Con ésto, resulta que:

$$CV = \frac{s}{\overline{x}} = \frac{5.1}{171} = 0.02982.$$

Como el valor de coeficiente es muy bajo, los datos son homogéneos.

1.10. Medidas de forma

Hasta ahora, hemos analizado la localización y la dispersión de una distribución, pero necesitamos conocer más sobre el comportamiento de los datos. En esta sección, analizaremos las medidas de forma[2].

Las medidas de forma de una distribución se clasifican en dos grupos: medidas de asimetría y medidas de curtosis.

1.10.1. Asimetría

El *coeficiente de asimetría* de una variable mide el grado de asimetría de la distribución de sus datos en torno a su media. Es adimensional y se define como sigue:

$$As = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3/n}{s^3}.$$

Las *colas* de una variable están constituidas por los valores alejados de la media (valores extremos). Una variable es asimétrica si su cola a un lado es más larga que su cola al otro y simétrica si ambas colas son igual de largas.

- si $As > 0$, la distribución será asimétrica a la derecha. La cola a la derecha es más larga que la cola a la izquierda.
- si $As = 0$ la distribución será simétrica. Ambas colas son igual de largas.
- si $As < 0$ la distribución será asimétrica a la izquierda. La cola a la izquierda es más larga que la cola a la derecha.

[2] En la definición de las medidas de forma no hay unidad de criterios entre los especialistas, por lo que hay una amplia variedad.

1.10.2. Apuntamiento o curtosis

El *coeficiente de apuntamiento* o *curtosis* de una variable sirve para medir el grado de concentración de los valores que toma en torno a su media. Se elige como referencia una variable con distribución normal, de tal modo que para ella el coeficiente de apuntamiento es cero.

$$Ap = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^4/n}{s^4} - 3.$$

Según su apuntamiento, una variable puede ser:

- Leptocúrtica, si $Ap > 0$; es decir, es más apuntada que la normal. Los valores que toma la variable están muy concentrados en torno a su media y hay pocos valores extremos.
- Mesocúrtica, si $Ap = 0$; es decir, es tan apuntada como la normal.
- Platicúrtica, si $Ap < 0$; es decir, es menos apuntada que la normal. Hay muchos valores extremos, las colas de la variable son muy pesadas.

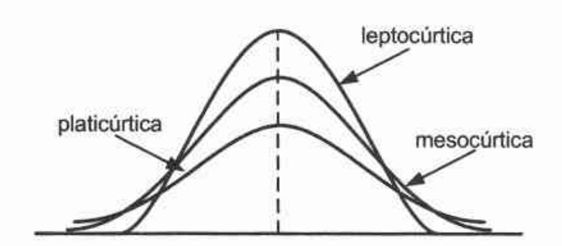


Figura 1.13: Curtosis de curvas simétricas.

Ejemplo. (Continuación.) Calcular los coeficientes de simetría y apuntamiento de los sueldos de diez personas que ganan (en dólares):

170 172 168 165 173 178 180 165 167 172.

Solución: Previamente se había calculado que $\overline{x} = 171$ y $s = 5.1$. Además,

$$\begin{aligned}\frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{n} &= \frac{(170-171)^3 + (172-171)^3 + \cdots + (167-171)^3 + (172-171)^3}{10} \\ &= 55.8.\end{aligned}$$

$$\begin{aligned}\frac{\sum_{i=1}^{n}(x_i - \overline{x})^4}{n} &= \frac{(170-171)^4 + (172-171)^4 + \cdots + (167-171)^4 + (172-171)^4}{10} \\ &= 1191.\end{aligned}$$

Entonces,

$$\begin{aligned} As &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3/n}{s^3} = \frac{55.8}{(5.1)^3} \\ &= 0.421. \end{aligned}$$

$$\begin{aligned} Ap &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})^4/n}{s^4} - 3 = \frac{1191}{(5.1)^4} - 3 \\ &= -1.239. \end{aligned}$$

Los datos son levemente asimétricos, con asimetría hacia la derecha; también, son platicúrticos, con posible presencia de valores atípicos.

1.11. Otras representaciones gráficas

Los gráficos analizados anteriormente no requieren realizar cálculos de medidas estadísticas. Los gráficos que a continuación se presentan, sí los emplean; por tanto, son más poderosos al realizar un análisis.

1.11.1. Diagrama de balanza

El *diagrama de balanza* fue introducido en el año 2000, como una herramienta que muestra, en un mismo gráfico, la forma de los datos, su valor central y su variabilidad al representar el promedio, el mínimo, el máximo y la desviación estándar de los datos.

Para su construcción se procede de la siguiente manera:

1. Se calcula el promedio, la desviación estándar, el mínimo y el máximo del conjunto de datos que se analiza.
2. Sobre una recta se ubican los valores del promedio, el mínimo y el máximo. Los segmentos que unen el promedio con el mínimo y con el máximo se denominan *brazos de la balanza.*
3. Sobre la misma recta se ubican dos puntos –uno a la izquierda y otro a la derecha de la media–, a una distancia igual a la desviación estándar.
4. Debajo del valor del promedio se dibuja un triángulo.

El diagrama queda así:

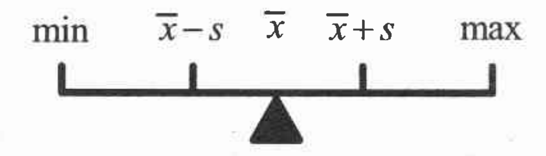


Figura 1.14:

El diagrama de balanza se interpreta de la siguiente manera:

1. Si los datos son simétricos, el valor del promedio se sitúa en el centro del gráfico.
2. Si los datos están agrupados en torno al centro, los brazos de la balanza serán cortos; por el contrario, si los datos están dispersos en torno al centro, los brazos de la balanza serán largos.
3. Si uno de los dos brazos de la balanza es mucho más largo que el otro, nos indica que los datos son asimétricos y que hay posible presencia de valores atípicos en las observaciones.

Puede ser útil combinar (sobre el mismo gráfico) con un diagrama de puntos para visualizar la manera en que se distribuyen las observaciones.

Ejemplo. Realizar el diagrama de balanza de los siguientes datos:

5	5	5	5	10	10	20	20	27	35
39	55	55	60	60	60	68	75	90	90

Solución: Estos datos tienen las siguientes características:

$$\text{mín} = 5, \quad \text{máx} = 90, \quad \overline{x} = 39.7, \quad s = 29.3.$$

Entonces,

$$\begin{aligned} \overline{x} - s &= 39.7 - 29.3 = 10.4. \\ \overline{x} + s &= 39.7 + 29.3 = 69.0. \end{aligned}$$

El diagrama de balanza se muestra a continuación:

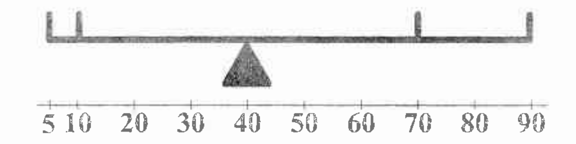


Figura 1.15:

Según se observa en el gráfico, el promedio no se encuentra en el centro del rango, entonces se deduce que los datos son asimétricos. Además, los brazos de la balanza no tienen igual longitud, lo que denota la posible presencia de valores atípicos en el extremo derecho.

1.11.2. Diagrama de caja

El *diagrama de caja* fue introducido en 1977, por John W. Tukey como una herramienta que muestra, en un mismo gráfico, la forma de los datos, su valor central y su variabilidad al representar la mediana, los cuartiles, el rango intercuartil y el rango de las observaciones.

Para su construcción se procede de la siguiente manera:

1. Sobre una línea horizontal se localizan la mediana, los cuartiles inferior y superior y los datos mínimo y máximo.

2. Se construye una caja angosta que une a Q_1 y Q_3; a continuación, se divide esta caja en dos mediante una línea que pase por Q_2.

3. Finalmente, se trazan las *vallas*, que son dos rectas, una desde cada extremo de la caja, hacia el valor mínimo y hacia el valor máximo de los datos.

En la Figura 1.16 se muestra un diagrama de caja.

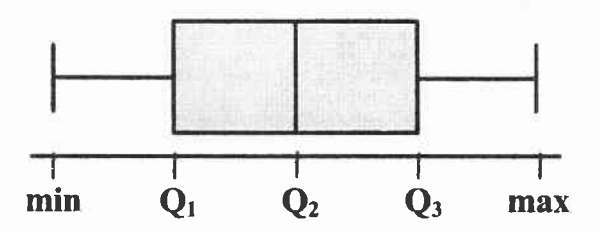


Figura 1.16: Diagrama de caja.

Un diagrama de caja es especialmente útil para examinar la simetría de los datos, la presencia de valores atípicos y para comparar dos conjuntos de muchos datos.

Ejemplos

1. **(Continuación.)** Trazar el diagrama de caja correspondiente a los datos de la estatura de 46 señoras, cuyas medidas son:

Estatura	1.45	1.48	1.50	1.53	1.55	1.57	1.60	1.63	1.65
Frecuencia	2	4	5	8	12	7	4	3	1

Solución: Antes se determinó que $Q_1 = 1.53$, $Q_2 = 1.55$, $Q_3 = 1.57$, $x_{\text{mín}} = 1.45$ y $x_{\text{máx}} = 1.65$.

El diagrama de caja es el siguiente:

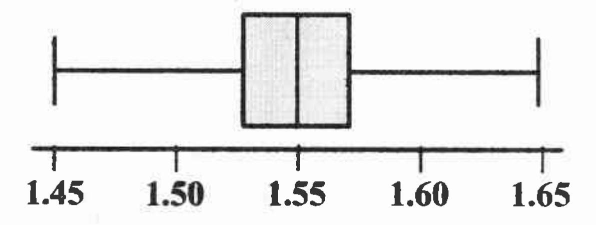


Figura 1.17:

Como se observa, los datos son bastante simétricos, con una fuerte concentración en torno al centro y –puesto que las vallas son largas– con la posible presencia de valores atípicos (el mínimo y el máximo).

2. Se recogieron los datos de los ingresos mensuales de 200 hombres y 250 mujeres, que realizan

trabajos no calificados, obteniéndose la siguiente tabla:

Ingreso	Hombres	Mujeres
180		10
190	5	55
200	20	75
210	35	25
220	47	40
230	75	45
240	20	

Comparar los ingresos de los dos grupos mediante sus diagramas de caja.

Solución: Se tiene la siguiente tabla que resume las medidas descriptivas requeridas:

	mín	Q_1	Q_2	Q_3	máx
Hombres	190	210	220	230	240
Mujeres	180	190	200	220	230

Con todos estos elementos, los diagramas de caja son

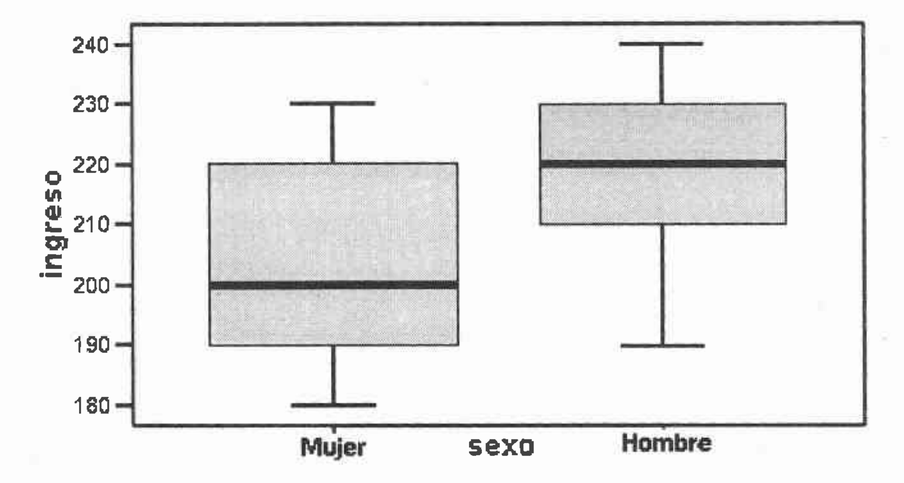


Figura 1.18:

En el diagrama correspondiente a las mujeres, observamos que la mediana no se encuentra en la mitad de la caja, denotando una asimetría, con fuerte concentración hacia valores bajos. Como las vallas son cortas, podemos inferir que no hay presencia de valores atípicos.

En el diagrama que corresponde a los hombres, se observa que la mediana está en la mitad de la caja, indicando que los datos son simétricos. Como la valla inferior es más larga que la superior, nos indica que un valor de 190 es atípico para los hombres.

De acuerdo a las posiciones de los diagramas, se observa que, en general, las mujeres tienen ingresos menores. También, se aprecia que los ingresos de los hombres están más concentrados alrededor de la mediana que los de las mujeres, denotando que aquellos son más homogéneos.

1.12. Ejercicios

1. Una persona está manejando un carro en una autopista a 70 km/h y nota que el número de autos a los que pasa es igual al número de autos que a ella le pasan. Los 70 km/h son el promedio, la mediana o la moda de las velocidades de los autos en la carretera. ¿Por qué?

2. Dadas $n = 8$ mediciones: 4, 2, 6, 5, 7, 5, 4, 6.

   Determine: a) $\overline{x}$; b) la mediana; c) s; d) el rango; e) la asimetría; f) la curtosis.

3. Dadas $n = 9$ mediciones: 5, 8, 8, 4, 4, 9, 7, 5, 4.

   Determine: a) $\overline{x}$; b) la mediana; c) s; d) el rango; e) el RIQ; f) la asimetría; f) curtosis.

4. En 1904, Cushny y Peebles publicaron en el artículo «The action of optimal isomers» (*Journal of Physiology*), un estudio sobre el efecto de dos isómeros de la molécula hidrocinamida hidrobromida en producir sueño. Se presentó la variación en el número de horas de sueño por noche al usar las dos versiones de la droga:

Paciente	Dextro	Levo
1	+0.7	+1.9
2	−1.6	+0.8
3	−0.2	+1.1
4	−1.2	+0.1
5	−0.1	−0.1
6	+3.4	+4.4
7	+3.7	+5.5
8	+0.8	+1.6
9	+0.0	+4.6
10	+2.0	+3.4

   a) Realice un diagrama de puntos para cada uno de los dos tipos de drogas y compárárelos. ¿Cuál de los dos isómeros es más efectivo en producir aumento en las horas de sueño?

   b) Realice un diagrama de tallo y hojas con los datos.

   c) Calcule el promedio, la mediana y la desviación estándar de los datos de las dos drogas. ¿Cuál es más efectiva? Explique.

5. Un inversor tiene ahorros repartidos en 3 depósitos con 2000, 5000 y 10 000 dólares, respectivamente. Si el primero le rinde un 5 % anual, el segundo un 4 % anual y el tercero un 2 % anual. ¿Cuál es el tipo de interés que recibe?

6. En una empresa se registró la edad (en años completos) de sus empleados, resultando la siguiente tabla:

31	49	36	39	56	29	57	41	40	51
45	61	40	39	47	27	36	37	16	37
51	18	29	34	42	38	62	31	28	25
36	40	46	37	49	25	21	39	35	37
56	35	48	44	42	43	49	22	25	28

   a) Determine el número de clases que se debe utilizar en la distribución de frecuencias;

   b) Construya la tabla de frecuencias y el histograma;

   c) ¿Qué porcentaje de los empleados es menor que 50?;

   d) ¿Qué porcentaje de los empleados es mayor que 35.5?

7. En una bodega de venta de licores se registró las principales características de 25 marcas de

whiskys:

No. de whisky	Precio de venta	Proporción de malta	Categoría	Tiempo de añejamiento	Nota de calidad
1	70	20	1	5	3
2	60	20	1	5	2
3	65	20	1	7.5	2
4	74	25	1	12	2
5	70	25	1	12	3
6	73	30	1	5	0
7	70	30	1	8	0
8	55	30	1	5	2
9	93	33	2	6.5	1
10	62	33	2	8	3
11	87	33	2	8.5	3
12	78	35	2	8.5	2
13	83	40	2	8	4
14	90	40	2	5.5	2
15	110	40	2	12	1
16	113	40	2	8.5	1
17	96	40	2	12	3
18	82	45	2	12	3
19	127	45	2	8.5	4
20	160	100	3	12	3
21	90	100	3	12	4
22	86	100	3	12	2
23	100	100	3	10	3
24	100	100	3	11	3
25	95	100	3	12	0

a) Identifique el tipo de dato que representa a cada una de las variables;

b) Realice un diagrama de tallo y hojas para el precio de venta y el tiempo de añejamiento;

c) Calcule el promedio, la moda y la mediana del precio, la proporción de malta y el tiempo de añejamiento;

d) Encuentre la desviación estándar, el RIQ y el coeficiente de variación del precio, la proporción de malta y el tiempo de añejamiento;

e) Calcule los coeficientes de asimetría y de apuntamiento del precio, la proporción de malta y el tiempo de añejamiento;

f) Realice un gráfico de barras de la categoría y de la nota de calidad.

8. Calcule el promedio, la mediana y la moda de las edades de 25 personas:

32	33	34	31	32	31	34	32	34	32	31	34	31
31	32	32	34	34	32	33	34	33	33	34	31	

9. Dados los datos y sus frecuencias:

x_i	2	5	7	10
n_i	8	12	16	14

Halle: a) $\overline{x}$; b) Mo; c) s; d) el rango.

10. Dados los datos y sus frecuencias:

x_i	1	3	6	9	10
n_i	8	20	25	10	2

Halle: a) Q_2; b) $\overline{x}$; c) s; d) RIQ.

11. Se realizó una investigación sobre el precio de zapatos deportivos, de similares características, en diversos almacenes de la ciudad, obteniéndose los siguientes datos (dólares):

50	43	39	43	40	38	35	25	37	32
49	43	39	44	40	38	33	26	36	30
49	43	39	44	40	38	33	27	36	30
47	41	39	45	40	37	33	27	35	30
46	41	38	46	40	37	32	28	35	28

a) Determine la distribución de frecuencias individuales de los datos;

b) Elabore la distribución de frecuencias con datos agrupados por clases;

c) A partir de la distribución obtenida, trace el histograma.

12. A continuación se dan los resultados de la estatura de 100 estudiantes:

Estatura (en cm)	155	160	165	170	175	180	185
No. de estudiantes	10	14	26	28	12	8	2

Halle:

a) la estatura promedio y la desviación estándar;

b) la media armónica y la media geométrica;

c) la mediana y el RIQ.

13. A partir de la siguiente distribución de frecuencias,

x_i	1.2	2.3	3.5	5.4	7.8	8.3	12.1
n_i	2	4	4	6	3	5	1

Encuentre:

a) los cuartiles inferior y superior y la mediana;

b) la media armónica;

c) la media geométrica.

14. La siguiente tabla muestra la temperatura nocturna (en °C) durante 200 días:

Intervalo	Frecuencia	Intervalo	Frecuencia
2 – 4	21	12 – 14	14
4 – 6	16	14 – 16	20
6 – 8	15	16 – 18	22
8 – 10	26	18 – 20	18
10 – 12	23	20 – 22	25

a) Determine: el promedio, la mediana y los cuartiles inferior y superior;

b) Construya el diagrama de caja de los datos.

15. Los siguientes datos se obtuvieron de una encuesta sobre las condiciones de vida, en el área rural de los cantones de Zapotillo y Macará y corresponden al número de hombres y de mujeres que

integran las familias encuestadas.

Hombres	Mujeres	Hombres	Mujeres	Hombres	Mujeres	Hombres	Mujeres
4	4	3	2	2	2	7	4
5	5	3	4	3	3	3	3
4	1	4	3	6	3	2	2
3	2	2	3	2	4	4	4
6	1	2	4	4	6	5	4
3	4	0	4	6	7	2	4
7	1	3	7	4	2	5	2
5	2	3	3	2	3	4	3
5	8	1	3	5	4	4	1

a) Realice un diagrama de puntos de los datos, clasificados por sexo;

b) Realice la tabla de frecuencias y el histograma de los datos, según el sexo de los encuestados;

c) Construya el diagrama de caja de los datos;

d) Interprete y compare los resultados obtenidos en a), b) y c);

e) Determine el número total de miembros en cada familia. Con estos nuevos datos trace el diagrama de puntos, el diagrama de tallo y hojas, la tabla de frecuencias, el histograma y el diagrama de caja. Interprete lo obtenido.

16. Las siguientes temperaturas fueron tomadas al medio día en Quito (en °C) durante una semana:

$$22, \quad 24, \quad 21, \quad 24, \quad 20, \quad 22, \quad 18.$$

a) Calcule el promedio y la desviación estándar de dichas temperaturas;

b) Para transformar los grados Celsius (c) en grados Fahrenheit (f) se usa la ecuación $f = 1.8c + 32$. Determine el promedio y la desviación estándar de las temperaturas en grados Fahrenheit;

c) Encuentre alguna relación entre los promedios y las varianzas calculados en a) y b).

17. En una investigación sobre la razón por la que frecuentemente habían colas muy largas en las cajas de un banco, se obtuvo información del tiempo (en minutos) requerido para atender a los clientes. Se tomaron 50 mediciones en una caja, las cuales se dan a continuación:

6.0	5.9	4.0	3.1	1.9	5.3	2.1	5.2	2.9	5.2
4.8	4.8	5.1	6.0	4.2	4.4	5.3	1.4	4.4	4.1
5.2	2.8	4.7	1.8	5.1	5.8	2.9	5.7	3.8	5.8
3.6	4.4	2.0	2.8	4.8	3.1	1.5	5.9	3.6	4.6
3.7	4.5	3.9	2.3	5.5	5.3	5.8	2.4	5.5	3.7

a) Calcule la desviación estándar y su aproximación a partir del rango;

b) Determine $(\overline{x} \pm s)$, $(\overline{x} \pm 2s)$ y $(\overline{x} \pm 3s)$;

c) Determine el número de observaciones que se encuentran en cada uno de los intervalos;

d) Construya el diagrama de caja de los datos y compare con los resultados de la parte b). ¿Qué observa?

18. La siguiente tabla muestra los tiempos de duración (en minutos) de las versiones en DVD de 22 películas dirigidas por Alfred Hitchcock:

Película	Tiempo	Película	Tiempo
The Birds	119	Dial M for Murder	105
Family Plot	120	Foreign Correspondent	120
Frenzy	116	I Confess	108
The Man Who Knew Too Much	120	Marnie	130
North by Northwest	136	Notorious	103
The Paradise Cane	116	Psycho	108
Rear Window	113	Rebecca	132
Rope	81	Shadow of a Doubt	108
Spellbound	111	Strangers on a Train	101
To Catch a Thief	103	Topaz	126
Under Capricorn	117	Vertigo	128

a) Construya un diagrama de tallo y hojas de los datos;

b) Calcule la mediana de los tiempos;

c) Calcule los cuartiles inferior y superior. Use esta información para detectar algún valor atípico y para trazar el diagrama de caja;

d) Determine el promedio y la desviación estándar;

e) Represente los datos mediante un diagrama de balanza. ¿Cuáles datos influyen más en los valores calculados?

f) Calcule los coeficientes de asimetría y de apuntamiento.

19. Las notas de un examen de 6 alumnos son: 6, 5, 9, 19, 3 y 18. Un alumno aprueba si su nota es mayor o igual que el promedio y que la mediana de las notas. ¿Qué porcentaje de los alumnos aprobaron el examen?

20. Un automóvil ha recorrido los 832 km que separan Loja de Esmeraldas, permutando regularmente las 5 llantas (incluida la de emergencia) para que todas tengan igual desgaste. ¿Cuál es el recorrido promedio de cada llanta?

21. El kilometraje que marca un auto, luego de 4 años de uso, es 100 mil kilómetros. Si el dueño lo compró nuevo y lo hace descansar 1 día, luego de usarlo 4 días seguidos, ¿cuál es el recorrido promedio diario de los días manejados, considerando años de 365 días?

22. De 400 alumnos de un colegio, cuya estatura promedio es 165 cm, 150 son mujeres y su estatura promedio es 160 cm. ¿Cuál es la estatura promedio de los varones?

23. Se tiene cuatro números. Al añadir el promedio de tres de ellos al número restante, se obtienen los números 17, 21, 23 y 29. Si se excluye al mayor de estos números, ¿cuál es el promedio de los tres restantes?

24. El promedio de 53 números es 600. Si se eliminan 3 números consecutivos, se observa que el nuevo promedio aumenta en 5 %. ¿Cuál es el mayor de dichos números consecutivos?

25. Calcule la mediana de las siguientes temperaturas:

Temp. (°C)	20.5	20.0	19.5	19.0	18.5	18.0	17.5
No. días	2	4	3	13	3	4	2

26. Calcule la mediana y la moda de los siguientes datos

Intervalo	Frecuencia
10 - 20	3
20 - 30	3
30 - 40	12
40 - 50	8
50 - 60	5

27. Los sueldos en una empresa son las siguientes:

1 gerente:	10 000
1 secretaria:	650
3 empleados:	500 (cada uno)
2 ayudantes:	400 (cada uno)
1 conserje:	300

¿Cuál es la medida de localización más representativa?

28. En una reunión hay 50 varones con una edad media de 20.5 años y 25 mujeres, las que en promedio son $\frac{1}{10}$ más jóvenes que los varones. Halle el número entero más próximo a la edad media de las personas de dicha reunión.

29. Un fumador dice que su vicio empezó con un cigarrillo en la primera semana, 2 en la segunda, 4 en la tercera, 8 en la cuarta, y así sucesivamente; hasta fumar casi 2 cajetillas diarias de 20 cigarrillos cada una, en promedio.

   a) ¿A cuántas semanas de haber empezado ocurrió ésto?;
   b) ¿Cuántos cigarrillos diarios, en promedio, fumó hasta la primera semana que llegó al máximo de su consumo?

30. Si cada uno de los 28 millones de habitantes de cierto país come, en promedio, 12 kg de pescado al año, entre conservas enlatadas y pescado fresco, siendo este rubro 4 veces el de conserva. ¿Cuántas toneladas de pescado fresco se consumen, en promedio, por año?

31. En una muestra de 20 empresas florícolas se obtuvieron los siguientes datos sobre el número de empleados y sus ingresos anuales, en miles de dólares:

No. de	Ingresos anuales		
empleados	50 – 100	100 – 250	250 – 1000
10 – 30	6	2	0
30 – 50	1	1	0
50 - 100	0	0	10

Calcule:

   a) el ingreso medio anual de las empresas;
   b) el número de empleados promedio.

32. De los datos de una tabla de distribución de frecuencias, con 5 intervalos de clase y ancho de clase común, se observó que: $Q_2 = 24$, $x_1 = 16$, $x_3 = 24$, $n_3 = 2n_1$, $n_5 = 2n_2$. ¿Qué porcentaje del total son menores de 30?

33. ¿A cuánto es igual la suma de cifras de la media aritmética de la siguiente serie de números?

$$\underbrace{11\ldots11}_{n\text{ cifras}},\ \underbrace{22\ldots22}_{n\text{ cifras}},\ \underbrace{33\ldots33}_{n\text{ cifras}},\ \ldots,\ \underbrace{99\ldots99}_{n\text{ cifras}}.$$

34. La siguiente tabla muestra la distribución de sueldos de 210 trabajadores de una empresa.

Sueldo	Trabajadores
600 – 700	100
700 – 800	20
800 – 900	60
900 – 1000	20
1000 – 1100	10

a) Halle la moda de los sueldos;

b) Debido al aumento de la productividad, los sueldos sufrieron un incremento del 10 % y, adicionalmente, un aumento de 50 dólares. Halle el nuevo sueldo promedio.

35. En una muestra de 1000 trabajadores, se registró sus sueldos en una tabla de frecuencias:

Sueldo	Trabajadores
0 – 400	150
400 – 800	300
800 – 1200	200
1200 – 1600	250
1600 – 2000	100

a) Calcule la moda de los datos;

b) ¿Qué porcentaje de los trabajadores tiene sueldos comprendidos entre el promedio y la mediana?

36. En la siguiente ojiva se muestran los sueldos de los trabajadores de un organismo estatal.

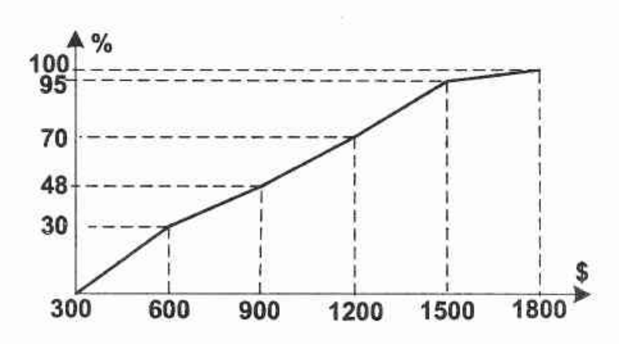


Figura 1.19:

Halle la diferencia entre el promedio y la mediana.

37. En la sección de pediatría de un hospital, los niños atendidos se clasificaron según su edad, obteniendo la siguiente tabla:

Edad	Frec. absoluta	Frec. absoluta acumulada	Frec. relativa	Frec. relativa acumulada
0 – 3			0.2	
3 – 6	20			
6 – 9				0.85
9 – 12		80		

Calcule el promedio, la mediana y la desviación estándar de la edad de los niños atendidos.

38. En la siguiente tabla se muestra la distribución de frecuencias de las ventas realizadas por los 60 locales de un centro comercial popular de la ciudad de Quito.

Ventas	Punto medio (x_i)	Frec. absoluta (n_i)	Frec. relativa	$n_i x_i$
–		12		
30 –			0.05	
–			0.30	
–				825
–	65			

Si los intervalos tienen igual longitud, halle el promedio, la mediana y la desviación estándar de las ventas.

39. En la siguiente tabla se muestra la distribución de frecuencias de los pesos de 100 personas:

Peso (kg)	Frec. absoluta	Frec. relativa	Frec. relativa acumulada
0 – 24		0.18	
24 – 48	26		
48 – 72			0.78
72 – 96			

Calcule la mediana del peso de estos individuos.

40. La siguiente tabla incompleta muestra la distribución de frecuencias de los depósitos bancarios realizados por 50 clientes, siendo el ancho de clase es constante e igual a 200.

Intervalo	Punto medio (x_i)	Frec. absoluta (n_i)	Frec. absoluta acumulada (N_i)	Frec. relativa (f_i)
–			9	
–				0.22
–	1100	12		
–				
–		7		
–				0.06

Luego de completar la tabla, calcule:

a) ¿cuántos clientes realizaron depósitos menores a 1000 dólares?;

b) ¿qué porcentaje de clientes realizaron depósitos entre 1200 y 1600 dólares?;

c) el promedio, la mediana y la moda de los depósitos.

Capítulo 2

El Concepto de Probabilidad

Las preguntas más importantes de la vida son, para la mayor parte, realmente solo problemas de probabilidad
Pierre Simon Laplace

En la naturaleza y en la vida cotidiana se presentan fenómenos cuyo resultado se determina anticipadamente mediante la aplicación de ciertas leyes o fórmulas; por ejemplo, los resultados de mediciones geométricas, los cálculos financieros o ciertos procesos físicos.

También existen fenómenos cuyo resultado no puede ser anticipado con certeza, sino que existe una *probabilidad* de que un cierto resultado se dé; por ejemplo, la ganancia que obtendrá un inversionista después de dos años, el tiempo que sobrevivirá un cónyuge a la muerte de su pareja o el número de autos que pasan por una esquina durante una hora determinada. Es evidente que nadie puede dar un resultado certero con anticipación a los tres *eventos* considerados, entonces si se da una respuesta, existe una incertidumbre en el resultado.

Para dar una explicación matemática a aquellos resultados que aparecen en experiencias en que está involucrado el azar, se desarrolló la teoría de probabilidades.

2.1. Reseña histórica

La presencia del hueso de astrágalo de oveja, que constituye el antecedente inmediato del dado, en las excavaciones arqueológicas más antiguas, parece confirmar que los juegos de azar tienen una antigüedad de más de 40 mil años. En la India, en el Rig-Veda (aproximadamente 1000 años a.C.), se menciona un juego de dados como un intento de medir la probabilidad. En Grecia, Sófocles atribuye a Palámedes la invención del juego de dados, durante el sitio de Troya. Así, en casi todas las culturas antiguas es posible encontrar referencias que nos indican que el estudio de los fenómenos aleatorios (dados, presencia de lluvia, el clima, etc.) fue muy importante.

En el Renacimiento se produjo un abandono progresivo de explicaciones teológicas, lo que condujo a una reconsideración de los experimentos de resultado incierto, y los matemáticos italianos del siglo XVI empezaron a interpretar los resultados de experimentos aleatorios simples. Por ejemplo, Cardano, en 1526, estableció, por condiciones de simetría, la equiprobabilidad de aparición de las caras de un dado. Por su parte, Galileo (1564 - 1642), respondiendo a un jugador que le preguntó por qué es más difícil obtener 9 tirando tres dados que obtener 10, razonó que de las 216 combinaciones posibles

equiprobables, 25 conducen a 9 y 27 a 10. Galileo publicó estos resultados en un tratado llamado *Considerazione sopra il giuoco dei dadi.*

El desarrollo del análisis matemático de los juegos de azar se produjo durante los siglos XVI y XVII. Algunos autores consideran como origen del cálculo de probabilidades la resolución del problema de los puntos en la correspondencia entre Pascal y Fermat en 1654. El problema planteado a estos autores por Chevalier de Meré, fue cómo debería repartirse el dinero de las apuestas, depositado en la mesa, si los jugadores se ven obligados a finalizar la partida sin que existiera un ganador. Aunque ninguno de estos dos matemáticos publicó al respecto, sí lo hizo Huygens en su tratado *Ratiociniis in ludo alae* (Razonamientos relativos al juego de dados). Su escrito tiene la trascendencia de ser el primer libro de probabilidades de la historia.

Durante el siglo XVIII, el cálculo de probabilidades se extendió a problemas físicos y de seguros marítimos. El factor principal de su desarrollo fue el conjunto de problemas de astronomía y de física que surgieron ligados a la constatación empírica de la teoría de Newton. Un primer problema fue el tratamiento de los errores de medición: se disponía de varias medidas independientes de una determinada magnitud física y se presentaba el interrogante de cómo combinarlas para obtener un resultado más preciso. Daniel Bernoulli (1700 - 1782) proporcionó la primera solución al problema de calcular una cantidad desconocida a partir de un conjunto de mediciones de su valor que, por el error experimental, presentan variabilidad.

Pierre Simón Laplace (1749 - 1827), introdujo la primera definición explícita de probabilidad y desarrolló la ley normal como modelo para describir la variabilidad de los errores de medida. En esta época también hubo importantes contribuciones de matemáticos como Legendre (1752 - 1833) y Gauss (1777 - 1855) para tratar de realizar predicciones del comportamiento de ciertos fenómenos.

Durante el siglo XIX, los matemáticos y astrónomos continuaron ampliando la teoría, de manera que a mediados de este siglo ya existían las herramientas que permitieron su consolidación como una rama científica. A pesar de ello, la aplicación de estos principios se restringía a la Física y la Astronomía.

Una descripción axiomática de la idea de probabilidad fue dada en 1933, por A. N. Kolmogorov. Ello constituyó la base de la moderna teoría, tal como hoy la conocemos. Con ésto, se consiguió elaborar modelos complejos y aplicar las probabilidades a muchas ciencias y campos de la vida.

En las últimas décadas, el empleo de la teoría de probabilidades en las modernas ciencias naturales, en las ciencias sociales y en ramas de aplicación, como la ingeniería, el cálculo actuarial o la economía, ha crecido enormemente y su conocimiento es una necesidad imprescindible.

Antes de iniciar el estudio de la probabilidad, revisemos los principales conceptos del análisis combinatorio.

2.2. Fundamentos de análisis combinatorio

Primero, definamos el *factorial* de un número entero positivo n como el producto

$$n! = n \times (n-1) \times \cdots \times 2 \times 1, \quad \text{con } 0! = 1.$$

Ahora, consideremos un conjunto finito compuesto por n elementos diferentes: $\{a_1, a_2, \ldots, a_n\}$. Se desea formar una colección constituida por k elementos ($k \leq n$). El número de estos subconjuntos depende de si los conjuntos son ordenados o no. Las colecciones ordenadas se llaman *variaciones* y las no ordenadas *combinaciones.*

Definición (de variación) Se denomina variación a cada uno de los arreglos ordenados de k elementos, tomados de otro de n elementos ($k \leq n$), de manera que estos arreglos difieren en algún elemento o en el orden de colocación.

El número de variaciones de k elementos que pueden obtenerse a partir de un conjunto de n elementos, denotado por $\mathbf{V}_n^k$, es igual a

$$\mathbf{V}_n^k = \frac{n!}{(n-k)!}.$$

Definición (de combinación) Se denomina combinación a cada uno de los subconjuntos de k elementos, tomados de otro de n elementos ($k \leq n$), sin tener en cuenta el orden de los mismos, de manera que no pueden haber dos combinaciones con los mismos elementos.

El número de combinaciones de k elementos que pueden obtenerse a partir de un conjunto de n elementos, denotado por $\mathbf{C}_n^k$, es igual a

$$\mathbf{C}_n^k = \frac{n!}{k!(n-k)!}.$$

A $\mathbf{C}_n^k$ se le denomina *coeficiente binomial*.

Ejemplo. Encontrar el número de variaciones y de combinaciones de dos elementos que se pueden obtener a partir del conjunto $\{a, b, c\}$.

Solución: Se tiene $n = 3$ y $k = 2$.

a) Se pueden formar $\mathbf{V}_3^2 = \frac{3!}{(3-2)!} = \frac{6}{1} = 6$ variaciones, que son:

$$(a,b),\ (b,a),\ (a,c),\ (c,a),\ (b,c),\ (c,b).$$

b) Se pueden formar $\mathbf{C}_3^2 = \frac{3!}{2!(3-2)!} = \frac{6}{2\cdot 1} = 3$ combinaciones:

$$\{a,b\},\ \{a,c\},\ \{b,c\}.$$

Definición (de permutación) Una permutación de n elementos es cada una de las variaciones de los n elementos distintos.

El número de permutaciones de n elementos es igual a

$$\mathbf{P}_n = n!.$$

Ejemplo. Encontrar las permutaciones que se pueden formar a partir del conjunto $\{a, b, c\}$.

Solución: Son $\mathbf{P}_3 = 3! = 6$ permutaciones; éstas son:

$$(a,b,c),\ (a,c,b),\ (c,a,b),\ (c,b,a),\ (b,c,a),\ (b,a,c). \quad ◄$$

Ahora, consideremos dos conjuntos de m y n elementos, respectivamente:

$$A = \{a_1, a_2, \ldots, a_m\} \quad \text{y} \quad B = \{b_1, b_2, \ldots, b_n\}.$$

Parejas. Con los m elementos de A y los n elementos de B es posible formar $m \times n$ parejas (a_j, b_k) que contengan un elemento de cada conjunto.

Ejemplo. En una fábrica de calzado se confeccionan 4 modelos de zapatos para damas, en 6 tamaños diferentes. Por lo tanto, se pueden fabricar $4 \times 6 = 24$ distintos tipos de zapatos.

Generalicemos este concepto a arreglos múltiples.

Arreglos múltiples. Consideremos los conjuntos $A = \{a_1, a_2, \ldots, a_m\}$ de m elementos, $B = \{b_1, b_2, \ldots, b_n\}$ de n elementos, y así sucesivamente hasta $G = \{g_1, g_2, \ldots, g_s\}$ de s elementos. Con ellos es posible formar $m \times n \times \cdots \times s$ arreglos $(a_i, b_j, \ldots, g_r)$ que contienen un elemento de cada conjunto.

Otra forma de ver este concepto es considerar un procedimiento A que se puede realizar de m maneras; un procedimiento B de n maneras; y así sucesivamente, hasta un procedimiento G de s maneras. La acción consistente en realizar el procedimiento A, seguido del procedimiento B, hasta llegar al procedimiento G; se puede efectuar de $m \times n \times \cdots \times s$ maneras diferentes.

Ejemplo. Suponga que se clasifica a un grupo de estudiantes universitarios según su sexo, estado civil y la carrera que estudian. El sexo puede ser masculino o femenino; el estado civil puede ser soltero, casado o divorciado; y, digamos que hay 7 carreras. Entonces, hay un total de $2 \times 3 \times 7 = 42$ clasificaciones diferentes. ◀

Anteriormente, se examinó las permutaciones de elementos de un conjunto, pero sin repetición; si ahora queremos determinar las permutaciones con repetición, bastará considerar en los arreglos múltiples el mismo conjunto.

Definición (de permutación con repetición) Una permutación con repetición, de k elementos obtenidos a partir de un conjunto de n elementos, es un arreglo de k elementos ordenados en el que los elementos pueden repetirse arbitrariamente.

El número de permutaciones con repetición es igual a

$$\mathbf{P}_n^k = n^k.$$

Ejemplo. Con los elementos del conjunto $A = \{a, b, c\}$, ¿cuántas permutaciones con repetición, de dos elementos, se pueden formar?

Solución: Se van a formar parejas considerando dos veces el conjunto A, por lo tanto se tiene $n = 3$ y $k = 2$; entonces, hay un total de $3^2 = 9$ permutaciones con repetición; ellas son:

$$(a, a),\ (a, b),\ (a, c),\ (b, a),\ (b, b),\ (b, c),\ (c, a),\ (c, b),\ (c, c).$$

2.3. Eventos y espacios muestrales

Examinemos un ejemplo: el lanzamiento de un dado una sola vez. Como resultado de la prueba se pueden producir diferentes resultados: «sale dos», «sale cinco», «el número que aparece es par», etc. Esto nos conduce a definir los *eventos.*

Definición (de evento) Se llama evento, notado como ω, a cualquiera de los resultados posibles de un experimento u otra situación que involucre incertidumbre.

Los eventos se clasifican en: *elementales*, aquellos que constan de un solo resultado; y *compuestos*, que consisten de más de un resultado. Por ejemplo, «sale dos» es un evento elemental; mientras

que «el número que aparece es par» es un evento compuesto, porque está conformado de los eventos elementales «sale dos», «sale cuatro» y «sale seis».

Observemos que todo evento relacionado con una prueba se puede describir en términos de eventos elementales.

Definición (de espacio muestral) La colección de todos los eventos elementales, notado por Ω, se denomina espacio muestral:

$$\Omega = \{\omega / \ \omega \text{ es evento elemental}\}.$$

Entonces, un evento no es más que un subconjunto del espacio muestral Ω.

Señalemos que el concepto de espacio muestral fue introducido por Galileo para resolver el problema de por qué en el lanzamiento de tres dados "10" y "11" aparecen más frecuentemente que "9" y "12". Para resolverlo listó todos los casos posibles.

Volviendo al ejemplo, si consideramos el número de puntos que aparecen al arrojar un dado, tenemos:

Espacio muestral: $\Omega = \{1, 2, 3, 4, 5, 6\}$.

$A = \{$ el número que sale es par $\} = \{2, 4, 6\}$.

Como los eventos se asocian a conjuntos, es natural pensar que sus operaciones tienen algún significado como eventos.

Sean A y B dos eventos de Ω, en el siguiente cuadro se presentan las equivalencias entre las proposiciones de las teorías de probabilidades y de conjuntos y en la Figura 2.1 se encuentran los diagramas de Venn correspondientes.

Notación	Interpretación en la teoría de conjuntos	Interpretación en la teoría de probabilidades
ω	Elemento o punto	Evento o suceso
Ω	Conjunto de puntos	Espacio muestral (suceso seguro)
$\emptyset$	Conjunto vacío	Evento imposible
$A \cup B$	Unión de conjuntos	Por lo menos uno de los eventos A o B ocurre
$A \cap B$	Intersección de conjuntos	Ambos eventos A y B ocurren
$A \setminus B$	Diferencia de conjuntos	A ocurre y B no ocurre
$A^c = \Omega \setminus A$	Conjunto complementario	No ocurre A
$A \cap B = \emptyset$	Conjuntos disjuntos	A y B se excluyen mutuamente (incompatibles)
$A \subseteq B$	A es subconjunto de B	Si A ocurre, también B

Es claro que estos conceptos se extienden a cualquier sucesión de eventos.

2.4. Definición axiomática de la probabilidad

Una probabilidad provee una descripción cuantitativa de la posibilidad de ocurrencia de un evento particular y se puede pensar que es su frecuencia relativa, en una serie larga de repeticiones de una prueba, en la que uno de los resultados es el evento de interés.

Formalmente, la probabilidad de un evento A se define como una función que cumple:

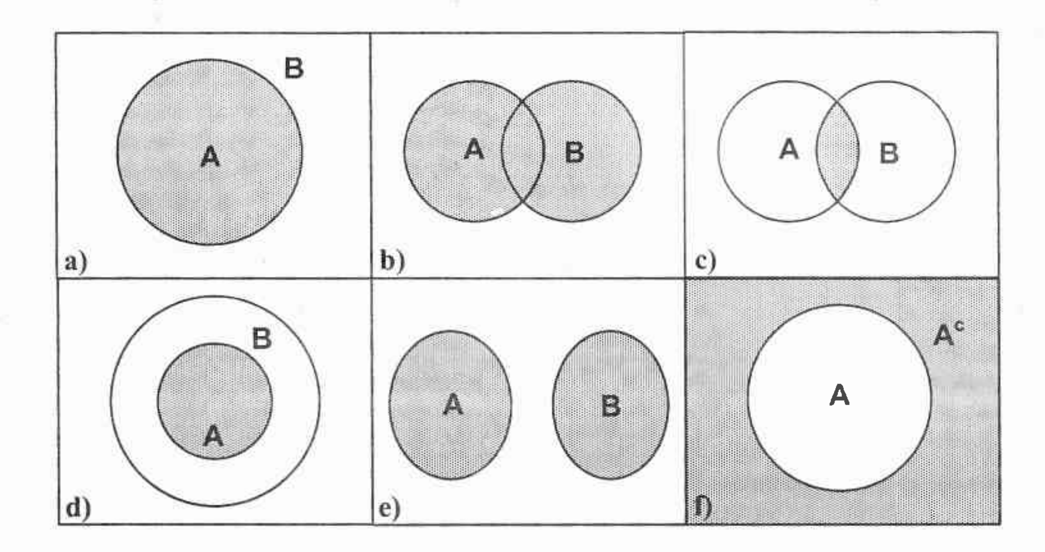


Figura 2.1: Interpretación de los conjuntos como eventos: a) Ocurre el evento A. b) Ocurre A u ocurre B $(A \cup B)$. c) Ocurre A y ocurre B $(A \cap B)$. d) Si A ocurre, también B $(A \subseteq B)$. e) Eventos incompatibles $(A \cap B = \emptyset)$. f) No ocurre A (ocurre A^c).

- A1. Para todo evento A: $0 \leq \Pr(A) \leq 1$.
- A2. $\Pr(\Omega) = 1$.
- A3. Si A y B son incompatibles: $\Pr(A \cup B) = \Pr(A) + \Pr(B)$.

De aquí, no es difícil demostrar que en general se cumple la relación:

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) \tag{2.1}$$

conocida como *fórmula de la probabilidad para la unión.*

Ejemplos

1. Dados los eventos A, B y C del espacio muestral Ω. Expresar mediante las operaciones entre conjuntos los eventos:

   a) Tan solo ocurre A.
   b) Si ocurre A, no ocurre B.
   c) Por lo menos dos de los eventos ocurren.

   Solución:

   a) Puede ocurrir A, y simultáneamente no ocurre B y no ocurre C; es decir que el evento es $x \in A \cap B^c \cap C^c$.
   b) Si no ocurre B entonces ocurre B^c; es decir que «si ocurre A, también ocurre B^c», el evento es $x \in A \subset B^c$.
   c) Ocurrirán (A y B) o (A y C) o (B y C) o (A y B y C), pero el último evento está contenido en los tres primeros. El resultado es: $x \in (A \cap B) \cup (A \cap C) \cup (B \cap C)$.

2. Demostrar que:

   a) $\Pr(A^c) = 1 - \Pr(A)$.

b) Si $A \subset B$ entonces $\Pr(A) \leq \Pr(B)$.

Solución:

a) $\Omega = A \cup A^c$ (con A y A^c disjuntos), entonces por A3., $\Pr(\Omega) = \Pr(A) + \Pr(A^c)$ y por A2., $\Pr(\Omega) = 1$; con lo que se obtiene: $1 = \Pr(A) + \Pr(A^c)$ y el resultado es inmediato.

b) Si $A \subset B$ entonces $B = A \cup (A^c \cap B)$ siendo A y $(A^c \cap B)$ incompatibles; por lo tanto, por A3. $\Pr(B) = \Pr(A) + \Pr(A^c \cap B)$.
Por A1., $\Pr(A^c \cap B) \geq 0$, entonces $\Pr(B) \geq \Pr(A)$. ◄

A continuación damos varias definiciones de mucha utilidad:

1. Dos eventos son *igualmente probables* si $\Pr(A) = \Pr(B)$.
2. El evento A es *más probable* que B si $\Pr(A) > \Pr(B)$.
3. *Evento cierto*.- Es el que siempre aparece en la realización de un experimento, su probabilidad es igual a 1.
4. *Evento imposible*.- Es aquel que jamás puede ocurrir, su probabilidad es igual a 0.

2.5. Cálculo de probabilidades

Al realizar el cálculo de las probabilidades es necesario distinguir de qué tipo de espacio muestral disponemos; ellos pueden ser: finito, infinito numerable o continuo.

2.5.1. Espacios muestrales finitos

Si consideramos el evento $A = \{\omega_1, \omega_2, \ldots, \omega_k\}$, su probabilidad está completamente determinada si conocemos sus valores en cada elemento $\Pr(\{\omega_1\})$, $\Pr(\{\omega_2\})$, ..., $\Pr(\{\omega_k\})$; entonces,

$$\Pr(A) = \sum_{i=1}^{k} \Pr(\{\omega_i\}). \tag{2.2}$$

Un caso particularmente importante se presenta cuando todas las probabilidades $\Pr(\omega)$ son iguales.

Si convenimos en designar $\text{Card}(A)$ el número k de elementos del conjunto A y $\text{Card}(\Omega)$ el número N de elementos del espacio muestral; entonces,

$$\begin{aligned} \Pr(A) &= \frac{\text{Casos favorables de } A}{\text{Casos posibles}} \\ &= \frac{\text{Card}(A)}{\text{Card}(\Omega)} = \frac{k}{N}. \end{aligned}$$

Es decir, la probabilidad de un evento aleatorio A es igual a la relación entre el número de eventos elementales favorables (cuando A sucede) y el número total de eventos elementales del espacio muestral. Esta definición es satisfactoria en problemas referentes a juegos de azar, loterías o experimentos sencillos.

En el ejemplo del lanzamiento de un dado consideremos el evento A «sale un número par»:

$$\Omega = \{1, 2, 3, 4, 5, 6\}, \qquad \mathrm{Card}(\Omega) = 6,$$
$$A = \{2, 4, 6\}, \qquad \mathrm{Card}(A) = 3.$$

$$\Pr(A) = \frac{\mathrm{Card}(A)}{\mathrm{Card}(\Omega)} = \frac{3}{6} = \frac{1}{2}.$$

En los siguientes ejemplos, consideraremos espacios muestrales finitos y aplicaremos los conceptos de análisis combinatorio al cálculo de probabilidades.

Ejemplos

1. En un estante hay 2 libros de historia y 3 de biología. Al azar, se toma un libro y luego se toma un segundo libro. Encontrar la probabilidad de que un libro de biología sea seleccionado: a) la primera vez; b) ambas veces.

   Solución:

   a) Por definición, $\Omega = \{H_1, H_2, B_1, B_2, B_3\}$.
   Sea A el evento «escoger un libro de biología»; es decir, $A = \{B_1, B_2, B_3\}$. Por tanto,

   $$\Pr(A) = \frac{\mathrm{Card}(A)}{\mathrm{Card}(\Omega)} = \frac{3}{5}.$$

   b) Que ambas veces se seleccione un libro de biología significa:

   - que la primera elección es un libro de biología, entonces se tiene 3 casos favorables; y
   - que la segunda elección también sea un libro de biología, entonces hay 2 casos favorables.

   Así, el número de casos favorables es igual a $3 \times 2 = 6$.
   El número de casos posibles, de todas las parejas sin repetición, es $5 \times 4 = 20$.
   Entonces, la probabilidad buscada es

   $$p = \frac{6}{20} = \frac{3}{10}.$$

2. En la final de un concurso escolar de matemática participan 6 alumnos, de los cuales 3 pertenecen al colegio A. Si se premia a los dos primeros con regalos diferentes, ¿cuál es la probabilidad de que los alumnos del colegio A obtengan los 2 premios?

   Solución: El conjunto Ω está constituido por las parejas que se pueden formar con los 6 participantes. El número total de parejas es $\mathbf{V}_6^2 = \frac{6!}{4!} = 30$.

   Sea el evento B: «ganan los alumnos del colegio A».

   El número de casos favorables en el cual 2 de los 3 alumnos del colegio A ganan los premios es: $\mathbf{V}_3^2 = 6$. Luego,

   $$\Pr(B) = \frac{6}{30} = 0.2.$$

3. Entre 100 fotografías de un sobre se encuentra la foto buscada. Del sobre se extraen al azar 10 fotos. Hallar la probabilidad de que entre ellas resulte la foto necesaria.

   Solución: El espacio muestral Ω está formado por los conjuntos de 10 elementos que pueden formarse a partir de 100: $\mathrm{Card}(\Omega) = \mathbf{C}_{100}^{10}$.

El número de resultados favorables que nos interesa es igual al total de formas como pueden escogerse 9 fotos de las 99 restantes; es decir, $\text{Card}(A) = \mathbf{C}_{99}^{9}$.

La probabilidad buscada es

$$\Pr(A) = \frac{\mathbf{C}_{99}^{9}}{\mathbf{C}_{100}^{10}} = \frac{1}{10}.$$

4. En el Consejo Universitario cada una de las 10 facultades está representada por el decano y el subdecano. Se nombra una comisión de 10 miembros elegidos al azar. Determinar la probabilidad de que:

   a) una determinada facultad esté representada;

   b) todas las facultades estén representadas.

Solución:

a) Considerando el evento complementario A^c: «una facultad dada no está representada», y calculemos su probabilidad. Hay 20 representantes, 18 de ellos no son de la facultad en cuestión, por lo tanto existen $\mathbf{C}_{18}^{10}$ casos favorables.

   El número de comisiones diferentes de 10 miembros que se pueden formar con los 20 miembros es $\mathbf{C}_{20}^{10}$, entonces

$$\Pr(A^c) = \frac{\mathbf{C}_{18}^{10}}{\mathbf{C}_{20}^{10}} = \frac{9}{38},$$

   finalmente,

$$\Pr(A) = 1 - \frac{9}{38} = \frac{29}{38} \approx 0.7632.$$

b) El número de maneras diferentes en que pueda estar un representante de cada facultad en la comisión es 2^{10}. La probabilidad del evento B: «todas las facultades están representadas» es

$$\Pr(B) = \frac{2^{10}}{\mathbf{C}_{20}^{10}} \approx 0.00554.$$

5. Se arrojan dos dados. Hallar la probabilidad del evento $A = \{$al menos en uno de los dos dados salen más de dos puntos$\}$.

Solución: El espacio muestral puede describirse como

$$\Omega = \{(i,j)/\, i,j = 1,2,\ldots,6\},$$

donde el evento elemental (i,j) corresponde a los i puntos aparecidos en un dado y los j puntos aparecidos en el otro. Consecuentemente, $\text{Card}(\Omega) = 36$.

Designemos como B_1 el evento consistente en que en el primer dado salen más de dos puntos y con B_2 el evento análogo para el segundo dado:

$$\begin{aligned} B_1 &= \{(i,j)/\, i = 3,4,5,6;\; j = 1,2,\ldots,6\}, \\ B_2 &= \{(i,j)/\, i = 1,2,\ldots,6;\; j = 3,4,5,6\} \end{aligned}$$

Por lo tanto, $\text{Card}(B_1) = \text{Card}(B_2) = 24$. Puesto que $B_1 \cap B_2 = \{(i,j)/\, i,j = 3,4,5,6\}$, entonces $\text{Card}(B_1 \cap B_2) = 4^2 = 16$. Ahora bien,

$$\Pr(B_1) = \Pr(B_2) = \frac{24}{36} = \frac{2}{3}, \qquad \text{y} \qquad \Pr(B_1 \cap B_2) = \frac{16}{36} = \frac{4}{9}.$$

De la fórmula de probabilidad para la unión se obtiene:

$$\begin{aligned}\Pr(A) &= \Pr(B_1 \cup B_2) = \Pr(B_1) + \Pr(B_2) - \Pr(B_1 \cap B_2) \\ &= \frac{2}{3} + \frac{2}{3} - \frac{4}{9} = \frac{8}{9}.\end{aligned}$$

Se recomienda que el lector resuelva este ejercicio mediante el empleo del evento complementario.

2.5.2. Espacios muestrales infinitos numerables

Sea $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n, \ldots\}$ un espacio muestral infinito numerable; entonces, resulta que

$$\sum_{i=1}^{\infty} \Pr(\{\omega_i\}) = 1,$$

luego, si A es un evento de Ω, su probabilidad se calcula por

$$\Pr(A) = \sum_{\omega_i \in A} \Pr(\{\omega_i\}).$$

Para el cálculo de las probabilidades, generalmente, se utilizan series numéricas infinitas.

Ejemplo. Juan y Andrés juegan tenis con la misma habilidad. Deciden jugar una secuencia de sets hasta que uno de ellos gane 2 sets seguidos. Halle la probabilidad de que se necesite jugar número par de sets para terminar el juego.

Solución: Sean los eventos: J: «gana el set Juan» y A: «gana el set Andrés».

Según el enunciado, el espacio muestral está conformado por los siguientes eventos elementales:

n	**Empieza Juan ganando**		**Empieza Andrés ganando**
1.	JJ	★	AA
2.	JAA		AJJ
3.	JAJJ	★	AJAA
4.	JAJAA		AJAJJ
5.	JAJAJJ	★	AJAJAA
	⋮		⋮

El evento B: «se jugará hasta que uno de ellos gane 2 sets consecutivos» es la unión de los eventos que están señalados con una estrella (★) en el espacio muestral.

Se tiene que

$$\begin{aligned}\Pr(\text{JJ}) + \Pr(\text{AA}) &= \frac{1}{2}, \\ \Pr(\text{JAJJ}) + \Pr(\text{AJAA}) &= \frac{1}{8}.\end{aligned}$$

Las restantes probabilidades son $\frac{1}{32}, \frac{1}{128}$, etc.

Entonces, la probabilidad de B está dada por la suma

$$\begin{aligned}\Pr(B) &= [\Pr(\text{JJ}) + \Pr(\text{AA})] + [\Pr(\text{JAJJ}) + \Pr(\text{AJAA})] + \cdots \\ &= \frac{1}{2} + \frac{1}{8} + \frac{1}{32} + \frac{1}{128} + \cdots + \frac{1}{2^{2n-1}} + \cdots\end{aligned}$$

La suma de esta serie geométrica es igual a $\frac{2}{3}$; por lo que $\Pr(B) = \frac{2}{3}$.

2.5.3. Espacios muestrales continuos

Supongamos que se tiene una figura plana Ω y dentro de ella se encuentra otra figura A (Figura 2.2). Sobre la figura Ω se ha marcado un punto al azar. Suponiendo que la probabilidad de que el punto caiga en A es proporcional al área de la figura y no de su forma o posición, la probabilidad de que el punto caiga en la figura A es:

$$\Pr(A) = \frac{\text{Área de } A}{\text{Área de } \Omega}.$$

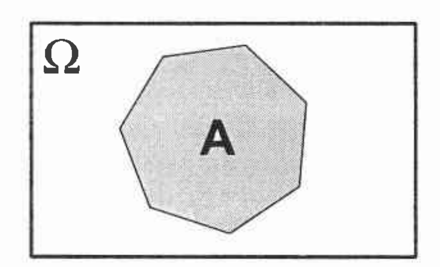


Figura 2.2: Interpretación geométrica de la probabilidad.

En general, si A es un evento de un espacio muestral continuo Ω, tal que su medida (longitud, volumen, tiempo, etc.) existe; entonces, su probabilidad es

$$\Pr(A) = \frac{\text{Medida de } A}{\text{Medida de } \Omega}.$$

Ejemplos

1. Sobre un plano se trazaron circunferencias concéntricas de radios 5 cm y 10 cm, respectivamente. Halle la probabilidad de que un punto marcado al azar en la circunferencia mayor caiga también en el anillo formado por las dos circunferencias (Figura 2.3).

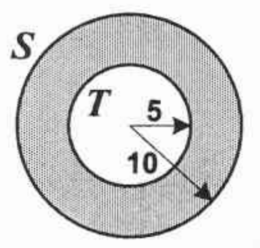


Figura 2.3:

Solución: El área del círculo mayor es $S = 10^2\pi\,\text{cm}^2 = 100\pi\,\text{cm}^2$.

El área del anillo comprendido entre las dos circunferencias es igual a la diferencia entre las dos áreas: $T = (10^2\pi - 5^2\pi)\,\text{cm}^2 = 75\pi\,\text{cm}^2$; entonces,

$$\Pr(A) = \frac{T}{S} = \frac{75\pi\,\text{cm}^2}{100\pi\,\text{cm}^2} = 0.75.$$

2. Sea $\Omega = \{(x, y)/\, 0 \leq x \leq 1;\ 0 \leq y \leq 1\}$ (Figura 2.4) el espacio muestral de un fenómeno aleatorio y suponiendo que todo punto de Ω tiene la misma probabilidad de ser tomado en cuenta.

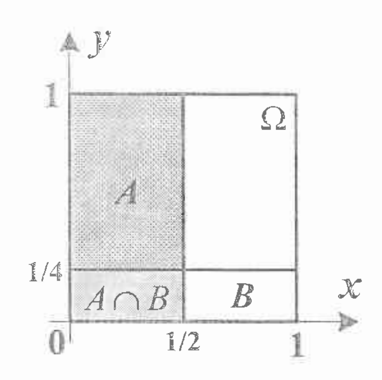


Figura 2.4:

Determinar la probabilidad de los eventos:

a) $A = \{(x, y)/\, 0 \leq x \leq 1/2;\ 0 \leq y \leq 1\}$;

b) $B = \{(x, y)/\, 0 \leq x \leq 1;\ 0 \leq y \leq 1/4\}$;

c) $A \cap B$;

d) $A \cup B$.

Solución:

a) Área de $\Omega = 1 \times 1 = 1$.
Área de $A = \frac{1}{2} \times 1 = \frac{1}{2}$, entonces $\Pr(A) = \frac{1/2}{1} = \frac{1}{2}$.

b) Área de $B = 1 \times \frac{1}{4} = \frac{1}{4}$, entonces $\Pr(B) = \frac{1/4}{1} = \frac{1}{4}$.

c) Área de $A \cap B = \frac{1}{2} \times \frac{1}{4} = \frac{1}{8}$, entonces $\Pr(A \cap B) = \frac{1}{8}$.

d) Por la fórmula de la unión,

$$\begin{aligned} \Pr(A \cup B) &= \Pr(A) + \Pr(B) - \Pr(A \cap B) \\ &= \frac{1}{2} + \frac{1}{4} - \frac{1}{8} = \frac{5}{8}. \end{aligned}$$

2.6. Ejercicios

Análisis combinatorio

1. Calcule los siguientes coeficientes binomiales C_n^k: a) C_4^2; b) C_3^0; c) C_3^1; d) C_3^2.

2. Calcule V_n^k en los siguientes casos: a) V_4^2; b) V_3^0; c) V_3^2; d) V_6^2.

3. Determine el número de permutaciones en un conjunto de n elementos para:
a) $n = 3$; b) $n = 4$; c) $n = 5$; d) $n = 6$.

4. Determine el número de parejas formadas por los elementos de los conjuntos A y B si:

a) $\text{Card}(A) = 4$, $\text{Card}(B) = 3$;
b) $\text{Card}(A) = 5$, $\text{Card}(B) = 4$;
c) $\text{Card}(A) = 8$, $\text{Card}(B) = 5$;
d) $\text{Card}(A) = 13$, $\text{Card}(B) = 5$.

5. Cuántos arreglos se pueden formar con los elementos de los conjuntos cuya cardinalidad se indica:
a) $\text{Card}(A) = 4$; $\text{Card}(B) = 2$; $\text{Card}(C) = 5$.
b) $\text{Card}(A) = 5$; $\text{Card}(B) = 7$; $\text{Card}(C) = 4$; $\text{Card}(D) = 5$.

6. Cuántas parejas con reposición pueden formarse con conjuntos cuya cardinalidad es:
a) $n = 3$; b) $n = 5$; c) $n = 7$; d) $n = 8$.

7. Forme todas las combinaciones y variaciones que se pueden obtener a partir de los conjuntos:

a) $A = \{a, e, i, o, u\}$ en grupos de tres elementos;
b) $B = \{1, 2, 3, 4, 5, 6\}$ en grupos de tres elementos.

8. Para los conjuntos indicados forme todas las parejas sin reposición y parejas con reposición:

a) $A = \{a, e, i, o, u\}$; b) $B = \{1, 2, 3, 4, 5, 6\}$.

9. Un comité de dirección de una empresa que consta de 4 gerentes y 6 subgerentes debe elegir un presidente y un vicepresidente. ¿De cuántas maneras se pueden elegir este par de funcionarios si el presidente debe ser un gerente?

10. Un hospital cuenta con 21 cirujanos con los cuales hay que formar ternas para realizar guardias. ¿Cuántas ternas se pueden formar?

11. Un amigo le quiere regalar a otro 3 discos y los quiere elegir entre los 10 que más le gustan. ¿De cuántas maneras puede hacerlo?

12. Al marcar un número telefónico una persona olvidó las tres últimas cifras, recordando que éstas son diferentes, las marcó al azar. Halle la probabilidad de que se haya marcado las cifras correctas.

13. De entre 9 empleados se deben seleccionar a 3 para viajar a 3 plantas A, B y C fuera de la ciudad. Cada empleado irá a una planta. ¿De cuántos modos se puede hacer la selección de los empleados que viajarán?

14. En el ejercicio anterior considérese que los 3 empleados van a ir a la misma planta. ¿De cuántas maneras se puede hacer la selección?

15. Si en el ejercicio anterior, de los 9 empleados, 7 son hombres. ¿Cuál es la probabilidad de seleccionar exactamente una mujer entre los tres escogidos?

16. ¿Cuántos números de 6 cifras pueden hacerse con los dígitos $\{1, 2, 3, 4, 5, 6\}$:

a) sin restricción alguna?; b) sin repetir ninguna cifra?; c) mayores que 500 000?

17. Siete personas han solicitado empleo para llenar dos vacantes. ¿De cuántos modos se pueden llenar las vacantes si:

a) la primera persona seleccionada recibe mayor salario que la segunda?;
b) no hay diferencia entre las vacantes?

18. ¿Cuántos partidos se juegan en un campeonato, en el que participan 20 equipos y en el que juegan todos contra todos, uno en casa y otro de visitante?

19. En un restaurante de comida rápida se indica al cliente que su hamburguesa, a más del pan y la carne, puede ir con todo lo siguiente o sin ello: salsa de tomate, mostaza, mayonesa, lechuga, cebolla, tomate o queso. ¿Cuántos tipos diferentes de hamburguesas son posibles?

20. La producción de una máquina consta de 4 fases. Hay 6 líneas de montaje para la primera fase, 3 para la segunda, 5 para la tercera, y 5 para la última. Determine de cuántas formas distintas se puede montar la máquina en este proceso de producción.

21. En un plano hay 15 puntos de los cuales no hay tres que sean colineales. ¿Cuántas rectas determinan?

22. ¿Cuántos triángulos determinan los vértices de un polígono regular de 9 lados?

23. Una heladería tiene 16 sabores disponibles. ¿De cuántas formas se pueden pedir 6 helados si:

    a) no se elije el mismo sabor más de una vez?;
    b) se puede pedir un mismo sabor hasta 6 veces?;
    c) un sabor no se puede pedir más de 5 veces?;
    d) la mitad debe ser de fresa?

24. Un entrenador de fútbol debe seleccionar a 11 jugadores de entre los que había convocado anteriormente para la concentración. Si puede hacer su selección de 12 376 maneras, ¿cuántos jugadores estuvieron presentes en la concentración? (Se supone que ningún jugador tiene un puesto fijo de juego.)

25. En un lenguaje de computación, un identificador consta de una letra o de una letra seguida de hasta siete símbolos, que pueden ser letras o dígitos. (En este lenguaje son indistinguibles las letras mayúsculas y minúsculas, hay 26 letras y 10 dígitos.) ¿Cuántos identificadores diferentes se pueden utilizar en el lenguaje de computación?

26. En cualquier set de un partido de tenis, el oponente X puede vencer al oponente Y de siete maneras. (Con el marcador $6-6$, se juega un desempate: tie breaker) El primer tenista que gane tres sets obtiene la victoria. ¿De cuántas maneras se pueden registrar los resultados si:

    a) X gana en cinco sets?;
    b) para ganar el partido se necesita jugar como mínimo tres sets?

27. ¿De cuántos modos se pueden poner 5 anillos diferentes en los dedos de una mano, omitiendo el pulgar?

Definición de probabilidad

28. Sean Ω un espacio muestral y A, B y C eventos cualesquiera, exprese las siguientes afirmaciones como uniones e intersecciones de A, B y C y de sus complementos.

    a) Ninguno de los eventos A, B, C ocurre;
    b) Por lo menos uno de los eventos A, B, C ocurre;
    c) No ocurre más que un evento;
    d) Ocurren exactamente dos eventos;
    e) Ocurren no más de dos eventos.

29. Con el empleo de la definición de probabilidad, demuestre:

a) $\Pr(\emptyset) = 0$;

b) $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$;

c) $\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$;

d) $\Pr(A) = \Pr(A \cap B) + \Pr(A \cap B^c)$.

30. Se arrojan dos dados, sean A el evento «la suma de las caras es impar», y B el evento «sale por lo menos un tres». Describa los eventos $A \cap B$, $A \cup B$, $A \cap B^c$. Encuentre sus probabilidades si se supone que los 36 eventos elementales tienen igual probabilidad.

31. Se consideran dos eventos A y B, tales que $\Pr(A) = \dfrac{1}{3}$ y $\Pr(B) = \dfrac{1}{2}$. Determine el valor de $\Pr(A^c \cap B)$ en los siguientes casos:

a) A y B son incompatibles;

b) $A \subset B$;

c) $\Pr(A \cap B) = \dfrac{1}{8}$.

32. Se consideran dos eventos A y B, con $\Pr(A) = 0.375$, $\Pr(B) = 0.5$ y $\Pr(A \cap B) = 0.125$. Calcule:

a) $\Pr(A^c)$ y $\Pr(B^c)$;

b) $\Pr(A \cup B)$;

c) $\Pr(A^c \cap B^c)$;

d) $\Pr(A^c \cap B)$ y $\Pr(A \cap B^c)$.

33. Sean A y B dos eventos tales que $\Pr(A) = 0.9$ y $\Pr(B) = 0.8$. Demuestre que $\Pr(A \cap B) \geq 0.7$.

34. Un experimento aleatorio consiste en arrojar una moneda y un dado a la vez y observar el resultado. Escriba el espacio muestral del experimento.

35. Una empresa tiene dos tiendas distribuidoras, una en el norte y otra en el sur de la ciudad. De los potenciales clientes, se sabe que el 30 % solo compra en la tienda norte, el 50 % solo compra en la tienda sur, el 10 % compra indistintamente en las dos tiendas y el 10 % de los consumidores no compra en ninguna de las dos. Sean los eventos A: «el cliente compra en la tienda norte» y B: «el cliente compra en la tienda sur». Calcule las probabilidades (e interprételas):

a) $\Pr(A)$;

b) $\Pr(A \cup B)$;

c) $\Pr(B^c)$;

d) $\Pr(A \cap B)$;

e) $\Pr(A \setminus B)$;

f) $\Pr(A^c \cap B^c)$;

g) $\Pr[(A \cap B)^c]$;

h) $\Pr(A \cup B^c)$.

36. En la intersección de una autopista, los automóviles pueden girar a la derecha (D) o a la izquierda (I). Desde un puesto de observación se registra el sentido de la maniobra de los tres primeros vehículos.

a) ¿Cuál es el espacio muestral del experimento?;

b) Sea A el suceso «a lo más uno de los coches gira a la derecha», B: «todos los vehículos giran en la misma dirección» y C: «exactamente uno de los coches gira a la derecha». ¿Qué relación existe entre los sucesos B y C?;

c) Enuncie y halle los elementos de los sucesos B^c, $B \cup C$, $A \cap B$, $A^c \cap B^c$.

Cálculo de probabilidades

37. Un gerente de compras desea hacer pedidos a proveedores diferentes, a los que nombra como A, B y C. Todos los proveedores son iguales en lo que respecta a la calidad por lo que escribe cada letra en un papel, mezcla los papeles y selecciona a ciegas a uno de ellos. Se hará el pedido al vendedor que salga seleccionado. Calcule las probabilidades de los eventos:

a) se seleccionó al proveedor B;

b) se selecciona al proveedor A o C;

c) el proveedor A no se selecciona.

38. Suponga que en un sorteo la probabilidad de ganar el primer premio es $\frac{2}{5}$ y la de ganar el segundo premio es $\frac{3}{8}$. Si la probabilidad de ganar al menos uno de los dos premios es $\frac{3}{4}$, ¿calcule la probabilidad de ganar solo uno de los dos premios?

39. Se envían 3 oficios a 3 personas diferentes. Sin embargo, una secretaria distraída revuelve los oficios y se puede considerar que los mandó al azar. Si una coincidencia es el hecho de que una persona reciba el oficio correcto, calcule la probabilidad de que haya:

    a) ninguna coincidencia;
    b) exactamente una coincidencia.

40. La fábrica ensambladora ha determinado que la demanda del auto Honda Civic es igual para cada uno de los colores azul, blanco, verde y rojo. Se hacen tres pedidos sucesivos de autos de ese modelo. Determine la probabilidad de que:

    a) se pidan uno azul, uno blanco y uno rojo;
    b) se pidan dos azules;
    c) se pida por lo menos uno verde;
    d) exactamente dos de los que se pidieron tengan el mismo color.

41. Luego de las pruebas para ocupar un puesto a los 6 aspirantes se les clasifica de acuerdo al puntaje obtenido. Los resultados no le llegan al empleador por lo que él contrata a dos aspirantes al azar. ¿Cuál es la probabilidad de que haya contratado a los dos aspirantes mejor calificados?

42. Un paquete de 6 focos tiene 2 unidades defectuosas. Si se escogen 3 focos para su uso, calcule la probabilidad de que ninguno tenga defectos.

43. En una caja hay 20 fotografías en la cual hay 6 mal tomadas. ¿Cuál es la probabilidad de seleccionar 2 fotografías defectuosas?

44. Entre 100 artículos de un lote hay 5 defectuosos. Halle la probabilidad de que entre 10 artículos escogidos al azar, no se tenga más de un artículo defectuoso.

45. Un distribuidor de electrodomésticos recibe un envío de 20 planchas, de las cuales hay 3 defectuosas. Para conocer si el lote está bueno prueba 6 aparatos. El distribuidor aceptará el lote si encuentra a lo más un aparato defectuoso entre los probados. ¿Cuál es la probabilidad de rechazar el envío?

46. De un ánfora, que contiene 100 boletos, se extraen tres boletos ganadores. ¿Cuál es la probabilidad de que gane una persona que compró:

    a) 4 boletos?;
    b) solo un boleto?

47. Entre las 80 estaciones de servicio que hay en una ciudad, 10 entregan una cantidad menor que la que el cliente compra. Un inspector de la Dirección de Hidrocarburos visita aleatoriamente cinco de ellas para verificar si la cantidad vendida es correcta. ¿Cuál es la probabilidad de que descubra al menos una fraudulenta?

48. En el juego del «cuarenta» se reparten 5 cartas, al azar, a cada jugador, a partir de un mazo de 40 cartas. ¿Cuál es la probabilidad de que un jugador tenga:

a) un as, un dos, un tres, un cuatro y un cinco, del mismo palo?;

b) 4 cartas del mismo palo?;

c) una «ronda»; es decir, 3 cartas de la misma denominación (as, dos, etc.)?

49. En un closet hay 6 pares de zapatos. Se escogen 4 zapatos al azar. Encuentre la probabilidad de que haya por lo menos un par de zapatos entre los 4 zapatos escogidos.

50. En los países europeos existe una forma muy popular de lotería, llamada Lotto, que consiste en seleccionar 6 números de una cartilla que contiene 44 números (del 1 al 44). El día del sorteo se seleccionan 6 bolas al azar y sin reposición. Una persona gana el premio principal si los 6 números sorteados coinciden con los seleccionados; también se puede ganar premios si 4 o 5 números sorteados coinciden. Determine la probabilidad de:

a) ganar el premio principal;

b) ganar al menos un premio.

51. Una persona presiona, al azar, 8 cifras en una calculadora. ¿Cuál es la probabilidad de los eventos siguientes:

a) A: «todas las cifras sean distintas»?;

b) B: «el producto de las 8 cifras es un número par»?;

c) C: «las 8 cifras forman un conjunto creciente»?;

d) D: «la suma de las cifras es igual a 3»?

52. En un círculo de 20 cm de radio se encuentra un círculo menor de radio 10 cm. Halle la probabilidad de que un punto marcado al azar en el círculo mayor caiga también en el círculo menor.

53. Dentro de un cancha de baloncesto, cuyas dimensiones son 20 m por 12 m, se encuentran dos charcos que tienen forma de círculos, de 8 y 5 m de diámetro respectivamente. ¿Cuál es la probabilidad de que una pelota lanzada a la cancha caiga dentro de uno de los charcos?

54. Dentro de un rectángulo de base 10 cm y altura 6 cm se encuentra un círculo que es tangente a 3 de los lados. Si se marca un punto al azar dentro del rectángulo, calcule la probabilidad de que el punto no se encuentre dentro del círculo.

55. Dentro del rectángulo limitado por las rectas $x = -\frac{\pi}{2}$, $x = \frac{\pi}{2}$, $y = -1$, $y = 1$, se tiene el gráfico de la función trigonométrica seno. Sobre el rectángulo cae una gota de tinta. ¿Cuál es la probabilidad de que la gota de tinta haya caído dentro del área comprendida entre el eje x y la curva $y = \operatorname{sen} x$? (Observación: Suponga que el área de la mancha de tinta es despreciable.)

2.7. Independencia y condicionalidad

En la teoría de probabilidad un concepto muy útil es el de independencia de eventos, que significa que la ocurrencia de uno de los eventos no da información sobre si otro evento ocurrirá o no; es decir, los eventos no influyen uno sobre otro.

Definición (de independencia) Dos eventos A y B se llaman independientes si la probabilidad de que ambos ocurran es igual al producto de las probabilidades de los dos eventos individuales. Es decir,

$$\Pr(A \cap B) = \Pr(A) \times \Pr(B).$$

Esta definición se puede extender a cualquier número de eventos.

Observación.- Si A y B son independientes, se puede demostrar que sus respectivos complementos son independientes; es decir, que se cumple:

- $\Pr(A \cap B^c) = \Pr(A) \times \Pr(B^c)$.
- $\Pr(A^c \cap B) = \Pr(A^c) \times \Pr(B)$.
- $\Pr(A^c \cap B^c) = \Pr(A^c) \times \Pr(B^c)$.

No se debe confundir los conceptos de eventos independientes y de mutuamente incompatibles (disjuntos).

Ejemplos

1. Sea $\Omega = [0,1] \times [0,1]$ y dados los eventos: $A = \{(x,y)/\, 0 \leq x \leq 1/2;\ 0 \leq y \leq 1\}$, $B = \{(x,y)/\, 0 \leq x \leq 1;\ 0 \leq y \leq 1/4\}$. Probar si A y B son independientes.

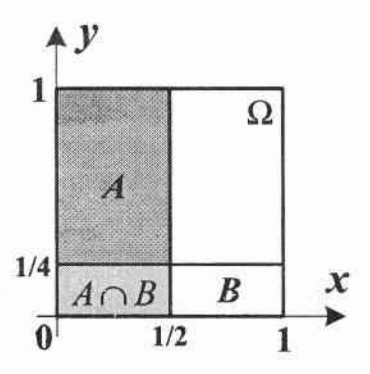


Figura 2.5:

Solución: Según la Figura 2.5 se tiene que $\Pr(A) = \frac{1}{2}$, $\Pr(B) = \frac{1}{4}$, entonces

$$\Pr(A) \times \Pr(B) = \frac{1}{2} \times \frac{1}{4} = \frac{1}{8}.$$

Por otro lado, antes se calculó que $\Pr(A \cap B) = \frac{1}{8}$.

Como se cumple que $\Pr(A \cap B) = \Pr(A) \times \Pr(B)$, entonces los eventos A y B son independientes.

2. En una máquina, para la señalización de emergencia se han instalado dos indicadores que funcionan independientemente. La probabilidad de que el indicador funcione durante una avería es de 0.95 para el primero y 0.9 para el segundo. Hallar la probabilidad de que durante una avería solo funcione un indicador.

Solución: Sean $A = \{\text{funciona el primer indicador}\}$ y $B = \{\text{funciona el segundo indicador}\}$.

El evento $C = \{\text{funciona solo un indicador}\}$ puede expresarse como $C = (A \cap B^c) \cup (A^c \cap B)$. Calculemos cada una de ellas:

$$\begin{aligned}
\Pr(A \cap B^c) &= \Pr(A) \times \Pr(B^c) = (0.95)(1 - 0.9) = 0.095. \\
\Pr(A^c \cap B) &= \Pr(A^c) \times \Pr(B) = (1 - 0.95)(0.9) = 0.045.
\end{aligned}$$

Por lo tanto,

$$\Pr(C) = \Pr(A \cap B^c) + \Pr(A^c \cap B) = 0.095 + 0.045 = 0.14.$$

3. Tres biólogos, independientemente uno del otro, midieron el contenido de suero en una muestra. La probabilidad de que cada uno cometa un error en la lectura del aparato es igual a 0.1, 0.15 y 0.2, respectivamente. Hallar la probabilidad de que en una sola medición por lo menos uno de los investigadores cometa un error.

   Solución: Sea el evento $A = \{$por lo menos uno de los investigadores comete un error$\}$, el complemento es $A^c = \{$ninguno de los investigadores comete un error$\}$.

   Calcularemos $\Pr(A^c)$, considerando que las mediciones son eventos independientes.

   Sean p_i la probabilidad de que el $i-$ésimo investigador cometa un error ($i = 1, 2, 3$), entonces

$$\begin{aligned} \Pr(A^c) &= (1 - p_1)(1 - p_2)(1 - p_3) \\ &= (1 - 0.1)(1 - 0.15)(1 - 0.2) = 0.612. \end{aligned}$$

   Resulta que $\Pr(A) = 1 - 0.612 = 0.388$. ◄

Un concepto estrechamente relacionado con la independencia es el de condicionalidad de eventos, que se lo puede enunciar de la siguiente manera: «se tiene fijo un cierto evento B, se desea conocer cuál es la probabilidad de que ocurra un evento A, sabiendo que ocurrió B».

Por ejemplo, suponga que usted va a almorzar al mismo lugar todos los viernes y que su almuerzo se sirve en 15 minutos (evento A) con probabilidad 0.9. Sin embargo, dado que usted nota que el restaurante está excepcionalmente lleno (evento B, fijo), la probabilidad de que sirvan su almuerzo en 15 minutos puede reducirse a 0.7. Ésta es la probabilidad condicional de ser servido en 15 minutos, dado que el restaurante está excepcionalmente lleno.

Definición (de probabilidad condicionada) Consideremos un espacio muestral Ω y un evento $B \in \Omega$ tal que $\Pr(B) \neq 0$. La probabilidad condicional de que un evento A ocurra, en el supuesto que B ha ocurrido, se representa por $\Pr(A|B)$ (que se lee «probabilidad de A, dado B»), se define como:

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

Ejemplos

1. En un estudio sociológico sobre la fidelidad en el matrimonio se obtuvo el siguiente modelo probabilístico, calificando al hombre y a la mujer como fiel (F) o infiel (I).

	Mujer	
Hombre	F	I
F	0.22	0.24
I	0.31	0.23

   a) ¿Cuál es la probabilidad condicional de que un esposo sea fiel, dado que su esposa es fiel?

   b) ¿Cuál es la probabilidad de que una esposa sea fiel, dado que su esposo es infiel?

   Solución: Convengamos en la siguiente notación de los eventos:

   HF: Hombre fiel, HI: Hombre infiel,

   MF: Mujer fiel, MI: Mujer infiel.

a) Deseamos calcular

$$\Pr(HF|MF) = \frac{\Pr(HF \cap MF)}{\Pr(MF)}.$$

De la tabla se obtiene que

$$\begin{aligned} \Pr(HF \cap MF) &= 0.22, \\ \Pr(MF) &= 0.22 + 0.31 = 0.53. \end{aligned}$$

Con ésto,

$$\Pr(HF|MF) = \frac{0.22}{0.53} = 0.415.$$

b) Calculemos

$$\Pr(MF|HI) = \frac{\Pr(MF \cap HI)}{\Pr(HI)},$$

con

$$\Pr(MF \cap HI) = 0.31 \quad \text{y} \quad \Pr(HI) = 0.31 + 0.23 = 0.54.$$

Entonces,

$$\Pr(MF|HI) = \frac{0.31}{0.54} = 0.574.$$

2. En un taller trabajan 7 hombres y 3 mujeres. Se escogen al azar 3 personas. Hallar la probabilidad de que todas las personas seleccionadas sean hombres.

Solución: Designemos los siguientes eventos:

A: el primer seleccionado es hombre,

B: el segundo seleccionado es hombre,

C: el tercer seleccionado es hombre.

- La probabilidad de que el primero sea hombre es $\Pr(A) = \frac{7}{10}$.
- La probabilidad de que el segundo sea hombre a condición de que el primero fue hombre es:

$$\Pr(B|A) = \frac{6}{9} = \frac{2}{3}.$$

- La probabilidad de que el tercero sea hombre sabiendo que los dos primeros también lo son, es la probabilidad de C dado A y B:

$$\Pr(C|A \cap B) = \frac{5}{8}.$$

La probabilidad buscada de que las tres personas escogidas sean hombres es

$$\Pr(A \cap B \cap C) = \Pr(A) \times \Pr(B|A) \times \Pr(C|A \cap B) = \frac{7}{10} \times \frac{2}{3} \times \frac{5}{8} = \frac{7}{24}.$$

2.8. Probabilidad completa y fórmula de Bayes

La probabilidad de un evento A, que puede ocurrir solo al aparecer uno de los eventos mutuamente excluyentes B_1, B_2, ..., B_n (Figura 2.6), tales que su unión es el espacio muestral, está dada por

$$\begin{aligned} \Pr(A) &= \Pr(B_1)\Pr(A|B_1) + \Pr(B_2)\Pr(A|B_2) + \cdots + \Pr(B_n)\Pr(A|B_n) \qquad (2.3) \\ &= \sum_{i=1}^{n} \Pr(B_i)\Pr(A|B_i). \end{aligned}$$

donde $\Pr(B_1) + \Pr(B_2) + \cdots + \Pr(B_n) = 1$.

La igualdad (2.3) se denomina la *fórmula de la probabilidad completa.*

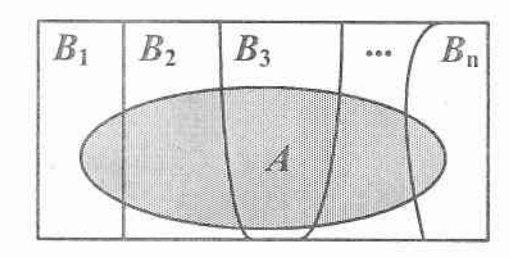


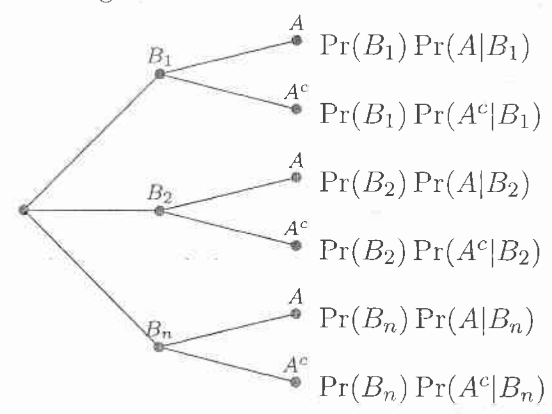
Figura 2.6: Partición del espacio muestral Ω.

Supongamos que el evento A puede ocurrir a condición de que aparezca uno de los eventos B_1, B_2, ..., B_n. Si A ya ocurrió, la probabilidad (condicional) del evento B_k es igual a

$$\Pr(B_k|A) = \frac{\Pr(A \cap B_k)}{\Pr(A)} = \frac{\Pr(B_k)\Pr(A|B_k)}{\sum_{i=1}^{n} \Pr(B_i)\Pr(A|B_i)}.$$

Esta igualdad se denomina *fórmula de Bayes.*

Para el cálculo mediante la fórmula de Bayes puede resultar conveniente disponer las probabilidades en un *diagrama de árbol* como el siguiente:



Esta disposición de los datos facilita la realización de los cálculos ya que únicamente se debe realizar una suma de los resultados en las ramas de interés.

Ejemplos

1. En una oficina hay 6 computadoras de marca y 4 clones. La probabilidad de que al utilizar una máquina, ésta encienda correctamente es 0.95 para las de marca y 0.8 para las clones. Un empleado utiliza al azar una computadora, hallar la probabilidad de que se encienda correctamente.

   Solución: Definamos los eventos:

   A: el empleado utiliza una máquina de marca,

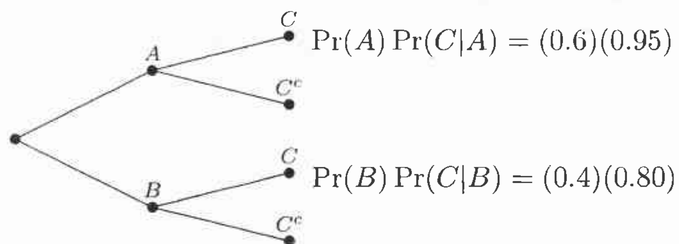
   B: el empleado utiliza una máquina clón,

   C: la máquina enciende correctamente.

Se tiene,

$$\Pr(A) = \frac{6}{10} = 0.6, \qquad \Pr(B) = \frac{4}{10} = 0.4,$$
$$\Pr(C|A) = 0.95, \qquad \Pr(C|B) = 0.8.$$

Si representamos las probabilidades en un diagrama de árbol se tiene:



Por la fórmula de la probabilidad completa,

$$\begin{aligned}\Pr(C) &= \Pr(A)\Pr(C|A) + \Pr(B)\Pr(C|B) \\ &= 0.6 \times 0.95 + 0.4 \times 0.8 = 0.89.\end{aligned}$$

2. Dos máquinas envasan gaseosa de manera automática, resultando que la primera envasa el doble que la segunda. La primera máquina envasa el 60 % de las botellas con la cantidad exacta y la segunda el 84 %. Una botella tomada del transportador resultó llena con la cantidad exacta. Hallar la probabilidad de que haya sido envasada por:

a) la primera máquina; b) la segunda máquina.

Solución: Designemos por eventos:

A: la botella está llena con la cantidad exacta;

B_1: la botella ha sido envasada por la primera máquina;

B_2: la botella ha sido envasada por la segunda máquina.

a) Se tiene

$$\Pr(B_1) = \frac{2}{3}, \qquad \Pr(B_2) = \frac{1}{3}.$$

La probabilidad condicional de que la botella contenga la cantidad exacta, si ha sido envasada por la primera máquina es

$$\Pr(A|B_1) = 0.6.$$

La probabilidad de que la botella contenga la cantidad exacta, si ha sido envasada por la segunda máquina, es

$$\Pr(A|B_2) = 0.84.$$

Por tanto, la probabilidad de que la botella tomada al azar contenga la cantidad exacta es

$$\begin{aligned}\Pr(A) &= \Pr(B_1)\Pr(A|B_1) + \Pr(B_2)\Pr(A|B_2) \\ &= \frac{2}{3} \times 0.6 + \frac{1}{3} \times 0.84 = 0.68.\end{aligned}$$

La probabilidad del evento «se escogió una botella con la cantidad exacta llenada por la primera máquina» es igual a

$$\Pr(B_1|A) = \frac{\Pr(B_1)\Pr(A|B_1)}{\Pr(A)} = \frac{\frac{2}{3} \times 0.6}{0.68} = \frac{10}{17}.$$

b) La probabilidad del evento «se escogió una botella con la cantidad exacta llenada por la segunda máquina» es

$$\Pr(B_2|A) = \frac{\Pr(B_2)\Pr(A|B_2)}{\Pr(A)} = \frac{\frac{1}{3} \times 0.84}{0.68} = \frac{7}{17}.$$

Este resultado también se puede calcular empleando el concepto de evento complementario.

3. En una ciudad, el 25 % de los habitantes son ancianos, el 35 % adultos y el 40 % son niños. Se sabe que la gripe afecta al 5 % de los ancianos, al 4 % de los adultos y al 2 % de los niños.

   a) Calcular la probabilidad de que un habitante, seleccionado aleatoriamente, tenga gripe.

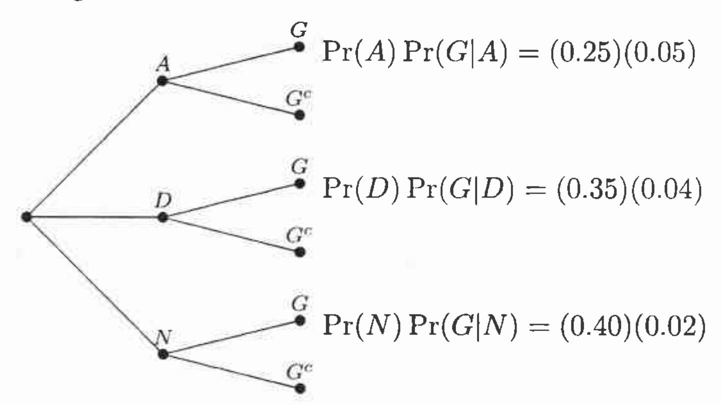
   b) Si un habitante tiene gripe, ¿cuál es la probabilidad de que éste sea anciano o niño?

Solución: Designemos los eventos:

A: el habitante es anciano. D: el habitante es adulto.

N: el habitante es niño. G: la persona tiene gripe.

a) Si utilizamos el diagrama de árbol tenemos:



Ahora, basta sumar los resultados parciales en las ramas para obtener el resultado deseado:

$$\begin{aligned} \Pr(G) &= \Pr(A)\Pr(G|A) + \Pr(D)\Pr(G|D) + \Pr(N)\Pr(G|N) \\ &= 0.25 \times 0.05 + 0.35 \times 0.04 + 0.4 \times 0.02 = 0.0345. \end{aligned}$$

La probabilidad de que un habitante de la ciudad tenga gripe es del 3.45 %.

b) Por la fórmula de Bayes:

$$\begin{aligned} \Pr(A|G) &= \frac{\Pr(A)\Pr(G|A)}{\Pr(G)} = \frac{0.25 \times 0.05}{0.0345} = \frac{125}{345}, \\ \Pr(N|G) &= \frac{\Pr(N)\Pr(G|N)}{\Pr(G)} = \frac{0.40 \times 0.02}{0.0345} = \frac{80}{345}. \end{aligned}$$

Consecuentemente,

$$\begin{aligned} \Pr(A \cup N|G) &= \Pr(A|G) + \Pr(N|G) \\ &= \frac{125}{345} + \frac{80}{345} = 0.594. \end{aligned}$$

La probabilidad de que si un habitante tiene gripe, éste sea anciano o niño, es del 59.4 %.

4. El 35 % de los créditos que otorga un banco es para vivienda, el 50 % para producción y el resto para consumo. Resultan morosos el 20 % de los créditos para vivienda, el 15 % de los créditos para producción y el 70 % de los créditos para consumo.

   a) Determine la probabilidad de que un crédito elegido al azar, se pague a tiempo.
   b) La probabilidad de que un crédito que ha resultado en mora, haya sido otorgado para la producción.

Solución: Designemos los eventos:

V: el crédito es para vivienda. P: el crédito es para producción.

C: el crédito es para consumo. M: el crédito está en mora.

a) Tengamos presente que el evento «el crédito se paga a tiempo» es el complemento del evento «el crédito está en mora»; entonces, buscamos $\Pr(M^c)$.
Por la fórmula de la probabilidad total,

$$\begin{aligned}\Pr(M) &= \Pr(M|V)\Pr(V)+\Pr(M|P)\Pr(P)+\Pr(M|C)\Pr(C)\\ &= 0.2\times 0.35+0.15\times 0.5+0.7\times 0.15=0.25.\end{aligned}$$

De manera que $\Pr(M^c)=1-\Pr(M)=1-0.25=0.75$.

b) Por la fórmula de Bayes,

$$\begin{aligned}\Pr(P|M) &= \frac{\Pr(P)\Pr(M|P)}{\Pr(M)}\\ &= \frac{0.5\times 0.15}{0.25}=0.3.\end{aligned}$$

2.9. Ejercicios

1. Sean A y B dos eventos con $\Pr(A)\neq 0$ y $\Pr(B)\neq 0$. Demuestre que

$$\Pr(A\cap B)=\Pr(B)\Pr(A|B)=\Pr(A)\Pr(B|A).$$

2. Demuestre que si A y B son eventos independientes y si $A\subseteq B$ entonces, $\Pr(B)=1$ o $\Pr(A)=0$.

3. Se consideran los eventos A y B tales que $\Pr(A)=\frac{1}{2}$; $\Pr(B)=\frac{1}{3}$; $\Pr(A\cap B)=\frac{1}{4}$. Calcule:

   a) $\Pr(A|B)$;
   b) $\Pr(B|A)$;
   c) $\Pr(A^c|B)$;
   d) $\Pr(B^c|A)$;
   e) $\Pr(A^c|B^c)$;
   f) $\Pr(B^c|A^c)$.

4. Suponga que un punto es elegido aleatoriamente en el cuadrado unitario. Si se conoce que el punto está en el rectángulo limitado por $y=0$, $y=1$, $x=0$ y $x=\frac{1}{2}$. ¿ Cuál es la probabilidad de que el punto esté en el triángulo limitado por $y=\frac{1}{2}$, $x=\frac{1}{2}$ y $y+x=\frac{1}{2}$?

5. Sea $\Omega=\{(x,y)/\,0\leq x\leq 1;\ 0\leq y\leq 1\}$ el espacio muestral de un fenómeno aleatorio. Calcule la probabilidad de los eventos:

   a) $A=\{(x,y)/\,0\leq x\leq 1;\ 0\leq y\leq 1/2\}$;
   b) $B=$ El triángulo limitado por las rectas $x=0$; $y=0$; $y=1-x$;

c) ¿Son independientes los eventos A y B?

6. En el cuadrado unidad se consideran los siguientes eventos:

A: El triángulo limitado por $x = 0$, $y = 1$, $y = x + 1/3$.

B: El triángulo limitado por $x = 0$, $y = 0$, $y = 1 - x$.

a) Halle $\Pr(B \setminus A)$, $\Pr(B|A)$ y $\Pr(A \cup B^c)$;

b) Pruebe si A y B son independientes

7. Un inspector debe seleccionar a un trabajador de entre 4 aspirantes numerados del 1 al 4. La selección la lleva a cabo mezclando los números y tomando uno al azar. Sean: A el evento «se seleciona al trabajador 1 o al 2»; B, el evento «se selecciona al trabajador 1 o al 3»; y C, el evento «se selecciona el trabajador 1». ¿Son independientes: a) A y B?; b) A y C?

8. Se lanzan dos dados, ¿cuál es la probabilidad de que en los dos dados salga el 3, si se sabe que la suma es 6?

9. En una biblioteca hay 8 libros de literatura de ciencia ficción, 3 de los cuales son de Isaac Asimov. La bibliotecaria toma al azar 2 libros. Determine la probabilidad de que ambos libros resulten ser de Isaac Asimov.

10. La Empresa de Correos ha determinado que el 10 % de los paquetes enviados al exterior no llegan a su destino. Dos libros se pueden enviar separadamente o en un solo paquete. Para cada una de las dos formas de envío postal, encuentre:

a) la probabilidad de que ambos libros lleguen a su destino;

b) la probabilidad de que al menos un libro llegue a su destino.

11. Suponga que el 5 % de todos los hombres y el 0.25 % de todas las mujeres sufren daltonismo. Una persona escogida al azar resulta ser daltónica. ¿Cuál es la probabilidad de que esta persona sea un hombre? (se considera que la cantidad de hombres y mujeres es igual).

12. El 35 % de los créditos de un banco es para vivienda, el 50 % para industrias y el 15 % para consumo. Resultan morosos el 20 % de los créditos para vivienda, el 15 % de los créditos para industrias y el 70 % de los créditos para consumo. Calcule la probabilidad de que se pague un crédito elegido al azar.

13. En una exhibición de arte hay 12 pinturas de las cuales 10 son originales. Un visitante selecciona una pintura al azar y decide comprarla después de escuchar la opinión de un experto sobre la autenticidad de la pintura. El experto está en lo correcto en 9 de cada 10 casos, en promedio.

a) Dado que el experto decide que la pintura es auténtica, ¿cuál es la probabilidad de que él no se equivoque?;

b) Si el experto decide que la pintura es una copia, entonces el visitante la devuelve y escoge otra, ¿cuál es la probabilidad de que la segunda pintura escogida sea original?

14. Hay una epidemia de cólera (C). Consideramos como uno de los síntomas la diarrea (D), pero este síntoma se presenta también en personas con intoxicación (I), e incluso en algunas que no tengan nada serio (N). Las probabilidades son:

$$\Pr(D|C) = 0.99; \quad \Pr(D|I) = 0.5; \quad \Pr(D|N) = 0.004$$

Se dan los siguientes porcentajes: el 2 % de la población tiene cólera y el 0.5 % intoxicación. Si una persona tiene diarrea calcule la probabilidad de que tenga cólera.

15. Una prueba para detectar el virus del SIDA en la sangre da el diagnóstico correcto con una probabilidad del 95 %. Según datos médicos, uno de cada 2 000 habitantes del país, en promedio, es portador del virus. Dado que la prueba fue positiva para una persona, ¿cuál es la probabilidad de que ella realmente tenga la enfermedad?

16. Una empresa financiera opera en las tres regiones del país: Costa, Sierra y Amazonía. El 50 % de las operaciones se realizan en la Costa, el 40 % en la Sierra y el resto en la Amazonía. Se ha estimado, debido a la larga experiencia, el porcentaje de clientes que no pagan sus deudas en cada una de las regiones. Para la Costa es del 1 %, para la Sierra del 2 % y para la Amazonía del 8 %. Si la empresa tiene 1000 clientes, determine cuántos pagan sus deudas puntualmente.

17. Una encuesta revela que el 70 % de la población tiene estudios secundarios, de los cuales el 12 % no tiene trabajo. Del 30 % que no tiene estudios secundarios, el 25 % no tiene trabajo. Calcule:

    a) El tanto por ciento de la población que no tiene trabajo;
    b) La probabilidad de que una persona elegida al azar tenga estudios secundarios entre las que no tienen trabajo.

18. De 200 aspirantes a un cargo se conoce la siguiente tabla respecto a experiencia en funciones similares y la formación académica necesaria

	Con formación	Sin formación
Con experiencia	16	32
Sin experiencia	24	128

Halle las probabilidades de encontrar una persona:

    a) con experiencia y con formación;
    b) con experiencia;
    c) con experiencia dado que tiene formación;
    d) sin formación dado que no tiene experiencia.

19. En una investigación sobre el crédito bancario a trabajadores agrícolas se obtuvo el siguiente modelo, en el que se califica al campesino como propietario o no propietario del terreno que cultiva y si mantiene o no mantiene deudas con los bancos.

	Propietario	
Deudor	SI	NO
SI	12	28
NO	20	64

Calcule la probabilidad de que:

    a) un campesino mantenga deudas con la banca;
    b) un campesino sea dueño del terreno que cultiva;
    c) un campesino sea propietario, dado que no es deudor;
    d) un campesino sea deudor, dado que es propietario del terreno.

20. A 100 empleados se les hizo un examen para determinar su destreza manual. Cuarenta de los empleados eran hombres. Sesenta de los empleados pasaron el examen porque alcanzaron una

calificación mayor que cierto nivel predeterminado de aprovechamiento. La clasificación entre hombres y mujeres fue la siguiente:

	Hombres (H)	Mujeres (M)
Pasaron (P)	24	36
No pasaron (N)	16	24

Suponga que se selecciona al azar un empleado de los 100 que hicieron el examen.

a) Calcule la probabilidad de que el empleado haya pasado y sea hombre;

b) Calcule la probabilidad de que el empleado sea hombre dado que pasó el examen;

c) ¿Son independientes P y H?;

d) ¿Son independientes P y M?

21. Los empleados de la compañía Cruz del Sur se encuentran distribuidos en 3 divisiones: Administración, Operación de Planta y Ventas. La siguiente tabla indica el número de empleados en cada división, clasificados por sexo.

	Mujeres (M)	Hombres (H)
Administración (A)	20	30
Operación (0)	60	140
Ventas (V)	100	50

Si se elige un empleado al azar,

a) ¿cuál es la probabilidad de que sea mujer?, ¿y de que trabaje en ventas?;

b) ¿cuál es la probabilidad de que sea hombre y trabaje en la división de Administración?;

c) ¿cuál es la probabilidad de que trabaje en la Operación de Planta si es mujer?;

d) ¿cuál es la probabilidad de que sea mujer si trabaja en la división de Ventas?;

e) ¿Son los sucesos V y H independientes? ¿y los sucesos A y M?

22. Dada la siguiente tabla que indica el comportamiento respecto del hábito de fumar en un grupo de 100 estudiantes que fueron averiguados.

	Hábito			
Sexo	No fuma	Fuma	Ex-fumador	TOTAL
Hombre	16	10	24	50
Mujer	30	16	4	50
TOTAL	46	26	28	100

a) Encuentre las distribuciones de las variables «sexo» y «hábito de fumar»;

b) Encuentre las probabilidades de los eventos: «la persona fuma» y «la persona fuma, dado que es mujer»;

c) ¿Son independientes los eventos «ser mujer» y «fumar»? ¿Por qué?

23. Del total de socios de un club, $\frac{3}{5}$ son hombres y $\frac{2}{3}$ son profesionales. Además, $\frac{1}{3}$ de las mujeres son no profesionales. Se elige al azar un miembro del club,

a) calcule la probabilidad de que sea hombre y profesional;

b) calcule la probabilidad de que sea hombre, dado que es profesional;

c) Determine si son independientes los eventos «ser mujer» y «no ser profesional».

24. En una fábrica, el 70 % de los empleados son lojanos. De entre los lojanos, el 50 % son hombres, mientras que de los no lojanos, sólo son hombres el 20 %.

    a) ¿Qué porcentaje de empleados no lojanos son mujeres?;
    b) Calcule la probabilidad de que un empleado de la oficina sea mujer;
    c) Fernando trabaja en dicha oficina. ¿Cuál es la probabilidad de que sea lojano?

25. En un país hay 4 partidos políticos que se dividen la opinión pública. Se sabe que:

    - El 35 % de la población adhiere al partido I.
    - El 31 % adhiere al partido II.
    - El 28 % adhiere al partido III.
    - El 6 % adhiere al partido IV.

    Entre los adherentes al partido I, un 36 % corresponde a personas con ingresos inferiores a dos salarios mínimos. Entre los adherentes al partido II, esa proporción es del 52 %. Para el partido III es un 42 %, y para el partido IV es 11 %. Si se elige una persona al azar y resulta tener un ingreso mayor a dos salarios mínimos, calcule la probabilidad que sea adherente al partido I.

26. La señora Sonia se fue de viaje y encargó a su hijo, Pablo, que riegue el rosal. La probabilidad de que Pablo olvide regar el rosal durante su ausencia es $\frac{2}{3}$. El rosal está en un estado inseguro: si se riega tiene igual probabilidad de secarse que de no secarse, pero solamente tiene un 0.25 de probabilidad de no secarse si no se riega. Después del viaje Sonia encuentra el rosal seco, ¿cuál es la probabilidad de que Pablo no lo haya regado?

27. Se estima que sólo un 20 % de los que compran acciones en Bolsa tienen conocimientos bursátiles. De ellos el 80 % obtienen beneficios. De los que compran acciones sin conocimientos bursátiles, sólo un 10 % obtienen beneficios. Se desea saber:

    a) El tanto por ciento de los que compran acciones en Bolsa que obtienen beneficios;
    b) Si se elige al azar una persona que ha comprado acciones en Bolsa y resulta que ha obtenido beneficios, ¿cuál es la probabilidad de que tenga conocimientos bursátiles?

28. En un supermercado el 70 % de las compras las realizan las mujeres; de las compras realizadas por estas, el 80 % supera los 20 dólares, mientras que de las compras realizadas por hombres sólo el 30 % supera esa cantidad.

    a) Elegido un comprobante de compra al azar, ¿cuál es la probabilidad de que supere los 20 dólares?;
    b) Si se sabe que el comprobante de compra no supera las 20 dólares, ¿cuál es la probabilidad de que la compra haya sido hecha por una mujer?

29. En una universidad existen tres facultades: A, B y C. En A hay matriculadas 150 chicas y 50 chicos; en B, 300 chicas y 200 chicos; y en C, 150 chicas y 150 chicos.

    a) Calcule la probabilidad de que un estudiante, elegido al azar, sea chico;
    b) Si un estudiante elegido al azar resultara ser chico, ¿cuál es su facultad más probable?

30. Entre los cinco aspirantes a un cargo de gerente, a dos se les considera excelentes y a los demás se les considera buenos. Para una entrevista se escoge al azar a dos de los cinco. Calcule la probabilidad de que se escoja:

a) a los dos excelentes;

b) por lo menos a uno de los excelentes;

c) a los dos excelentes dado que se sabe que por lo menos uno de los seleccionados es excelente.

31. Se dispone de dos métodos A y B para enseñar una destreza manual. El índice de reprobados es del 20 % para el método A y 10 % para el método B. Sin embargo, el método B es más caro por lo que solo se le usa el 30 % del tiempo y el A el otro 70 %. A un trabajador se le adiestra con uno de los dos métodos, pero no puede aprender en forma correcta. ¿Cuál es la probabilidad de que se le haya adiestrado con el método A?

32. En los exámenes de ingreso a una universidad cada candidato es admitido o rechazado de acuerdo a si él ha aprobado o reprobado la prueba. De los candidatos que realmente son capaces, el 80 % pasa la prueba; y de los que no son capaces, el 25 % pasan la prueba. Dado que el 40 % de los candidatos son realmente capaces, encuentre la proporción de estudiantes capaces que ingresan a la universidad.

33. Según datos de investigaciones genéticas se ha establecido que: los padres de ojos claros y los hijos de ojos claros constituyen el 5 % de las personas estudiadas; los padres de ojos claros y los hijos de ojos oscuros el 7.9 %; los padres de ojos oscuros y los hijos de ojos claros el 8.9 %; los padres de ojos oscuros y los hijos de ojos oscuros el 78.2 %. Halle la probabilidad de que:

a) el hijo sea de ojos oscuros, si el padre es de ojos oscuros;

b) el hijo sea de ojos claros, dado que el padre es de ojos claros.

34. Como un acto de buena vecindad Dios y Satanás acordaron un intercambio cultural entre el Cielo y el Infierno. Demonios del Infierno van a vivir en el Cielo, mientras que ángeles del Cielo van a vivir en el Infierno. Los demonios tienden a no decir la verdad más frecuentemente que los ángeles. Los demonios mienten el 80 % de las veces y los ángeles mienten el 20 % de las veces (¡en estos días es difícil encontrar ángeles buenos!). Después del intercambio, la proporción entre los demonios y ángeles en el Cielo es 2:3. Mi amigo José murió y fue al Cielo. Él encuentra a una persona en la calle y le pregunta donde encontrar un baño para hombres. Desafortunadamente, los demonios y los ángeles no se pueden distinguir por su aspecto físico. Deseamos determinar:

a) ¿Cuál es la probabilidad de que la respuesta haya sido una verdad a la pregunta de José?

b) Dado que la respuesta fue una mentira, ¿cuál es la probabilidad de que haya sido dada por un demonio?

35. Una compañía de tarjetas de crédito encuentra que cada mes el 50 % de quienes poseen la tarjeta cubren totalmente sus deudas.

a) Si se seleccionan dos usuarios al azar, ¿cuál es la probabilidad de que ambos paguen totalmente su deuda ese mes?;

b) Si se selecciona un cliente al azar, ¿cuál es la probabilidad de que dicha persona pague totalmente sus deudas en dos meses consecutivos?

c) ¿En qué hipótesis se apoyó para responder a los dos apartados anteriores? ¿Le parece que alguna de ellas no es razonable?;

d) Un examen más detallado de los registros de la compañía muestra que el 90 % de los clientes que pagan totalmente una cuenta mensual también lo hacen al mes siguiente y que sólo el 10 % de los que no pagan totalmente en un mes cubren totalmente su deuda al mes siguiente. Calcule, en este caso, la probabilidad pedida en b).

e) Con las hipótesis de d), calcule la probabilidad de que un cliente seleccionado al azar no pague totalmente ninguna de las dos cuentas mensuales consecutivas;

f) Calcule la probabilidad de que sólo pague una de las dos cuentas.

36. Basándose en varios estudios, una compañía ha clasificado, de acuerdo con la posibilidad de encontrar petróleo, las formaciones geológicas en 3 tipos. La compañía pretende perforar un pozo en un determinado lugar, al que le asignan las probabilidades de 0.35, 0.40 y 0.25 para los tres tipos de formaciones, respectivamente. De acuerdo con la experiencia, se sabe que el petróleo se encuentra en un 40 % de las formaciones de tipo I, en un 20 % de las de tipo II y en un 30 % de las de tipo III. Si tras perforar el pozo, la compañía descubre a su pesar que allí no había petróleo, determinar la probabilidad de que ese lugar se corresponda con una formación del tipo II.

37. El cardinal de un espacio muestral finito es m. Los eventos A y B son independientes y cumplen que:

$$\Pr(A) + \Pr(B) = p \quad \text{y} \quad \Pr(A \cap B) = \frac{p^2}{4}.$$

Halle la cardinalidad de A.

38. Demuestre que si se tienen B_1, B_2, ..., B_n eventos mutuamente excluyentes, tales que su unión es el espacio muestral, entonces se tiene que

$$\sum_{i=1}^{n} \Pr(B_i|A) = 1.$$

Capítulo 3

Variables Aleatorias, Esperanza y Varianza

Una variable aleatoria es el alma de una observación
Una observación es el nacimiento de una variable aleatoria
D. G. Watts, (1991)

En este capítulo introduciremos el concepto de variable aleatoria, que nos facilitará la realización del análisis de las principales características de los experimentos aleatorios y permitirá definir las leyes de probabilidad que ellos siguen, de manera muy general.

3.1. Variables aleatorias

El resultado de una prueba aleatoria no siempre es un número; por ejemplo, en el lanzamiento de una moneda los resultados son «cara» y «escudo». Sin embargo, a cada uno de los eventos le podemos asociar un número y sobre ellos aplicar las leyes que rigen la probabilidad.

Consideremos la definición clásica de función real, donde la cantidad y se llama función del número x, si a todo valor x de la variable independiente, le corresponde un valor y de la variable dependiente. Si esta idea la extendemos, se define una función donde la variable independiente no sea un número real sino que, en nuestro caso, sea un espacio muestral.

Definición (de variable aleatoria) Se llama variable aleatoria a cualquier función definida en un espacio muestral Ω con recorrido en un subconjunto finito o infinito de $\mathbf{R}$.

Es decir, la función

$$\begin{array}{rcl} X:\Omega & \longrightarrow & \mathbf{R} \\ \omega & \longmapsto & X(\omega) \end{array}$$

donde ω es un evento, es una variable aleatoria. Figura (3.1).

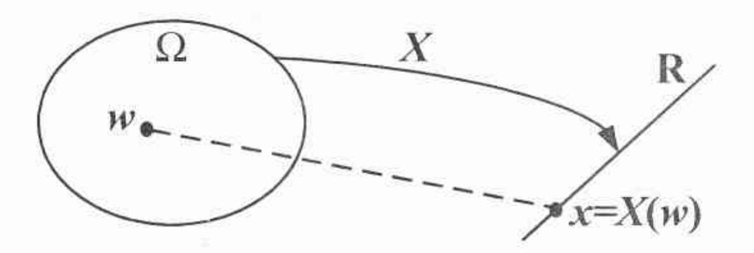


Figura 3.1:

Ya hemos estado trabajando con variables aleatorias sin referirnos explícitamente a ellas; por ejemplo, al arrojar un dado son posibles seis casos. Designando por ω_i el evento elemental consistente en salir i puntos, tenemos:

$$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}.$$

La variable aleatoria $X(\omega_i) = i$ identifica al número i de puntos obtenidos al lanzar el dado, se define así:

$$\begin{aligned} X(\omega_1) &= 1, \text{ corresponde al evento elemental } \{\text{aparece un punto}\}, \\ X(\omega_2) &= 2, \text{ corresponde al evento elemental } \{\text{aparecen dos puntos}\}, \\ X(\omega_3) &= 3, \text{ corresponde al evento elemental } \{\text{aparecen tres puntos}\}, \\ X(\omega_4) &= 4, \text{ corresponde al evento elemental } \{\text{aparecen cuatro puntos}\}, \\ X(\omega_5) &= 5, \text{ corresponde al evento elemental } \{\text{aparecen cinco puntos}\}, \\ X(\omega_6) &= 6, \text{ corresponde al evento elemental } \{\text{aparecen seis puntos}\}. \end{aligned}$$

Al arrojar una moneda tenemos dos eventos: C: «sale cara» o E: «sale escudo»; definimos la variable aleatoria X, que cuenta el número de caras aparecidas en una serie de lanzamientos, de la siguiente manera:

$$\begin{aligned} X(\omega) &= 1, \text{ si «sale cara»;} \\ X(\omega) &= 0, \text{ si «sale escudo».} \end{aligned}$$

La definición de la variable aleatoria depende del fenómeno a investigar.

Notación. Para evitar una escritura engorrosa, a los ω tales que $X(\omega) = t$ se los notará como $X = t$. La probabilidad $\Pr(\{\omega / X(\omega) = t\})$ se pondrá como $\Pr(X = t)$. De manera análoga se escribirá la $\Pr(\{\omega / X(\omega) \in (a, b]\})$ como $\Pr(a < X \leq b)$. Y así para cualquier intervalo convenientemente definido.

Las variables aleatorias se clasifican en *discretas* y *continuas*, de acuerdo a los valores que ellas tomen.

3.1.1. Variables aleatorias discretas

Definición (de variable aleatoria discreta) La variable aleatoria en cuyo recorrido el conjunto de los puntos que tienen probabilidad estrictamente positiva es finito o infinito numerable se llama variable aleatoria discreta.

Si el recorrido de la variable aleatoria X es el conjunto de números $\{x_1, x_2, \ldots x_n, \ldots\}$, se tiene que

$$\Pr\{X = x_i\} > 0.$$

Además, si $p_i = \Pr(X = x_i)$, es la probabilidad de que X tome el valor x_i, se cumple que

$$p_1 + p_2 + \cdots + p_n + \cdots = 1.$$

En otras palabras, X es discreta si una unidad de masa de probabilidad está distribuida sobre el eje real, concentrándose una masa positiva en cada punto de cierto conjunto finito o infinito numerable y en los restantes puntos no hay masa.

Variables aleatorias discretas son usualmente (pero no necesariamente) conteos de ciertos elementos; por ejemplo, el número de hijos de una familia, el número de ventas realizadas por un almacén, etc.

Una vez que se ha determinado las probabilidades asociadas a cada uno de los valores de una variable aleatoria discreta, es útil ponerlas en forma de una *distribución de probabilidad*, que es una tabla con todos sus posibles valores y sus correspondientes probabilidades, como la siguiente:

x	1	2	$\cdots$	n
$\Pr(X = x)$	p_1	p_2	$\cdots$	p_n

Ejemplos

1. Consideremos la prueba consistente en arrojar tres monedas. Tenemos que

$$\Omega = \{\{CCC\}, \{CCE\}, \{CEE\}, \{EEE\}\}.$$

Si X es la variable aleatoria que cuenta el número de escudos resultantes, X puede tomar los valores 0, 1, 2 y 3. De manera que

$$\begin{aligned} p_0 &= \Pr(X = 0) = \Pr(\{CCC\}) = \frac{1}{8}, \\ p_1 &= \Pr(X = 1) = \Pr(\{CCE\}) = \frac{3}{8}, \\ p_2 &= \Pr(X = 2) = \Pr(\{CEE\}) = \frac{3}{8}, \\ p_3 &= \Pr(X = 3) = \Pr(\{EEE\}) = \frac{1}{8}. \end{aligned}$$

Se tiene que su ley de probabilidad es

x	0	1	2	3
$\Pr(X = x)$	1/8	3/8	3/8	1/8

y se cumple que

$$p_0 + p_1 + p_2 + p_3 = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8} = 1.$$

2. Consideremos la siguiente prueba: se dispara con una pistola a un blanco situado a cierta distancia. Nos interesa analizar los eventos ω_i: «número de balas empleadas por un tirador hasta que se da en el blanco por primera vez».

Definimos la variable aleatoria X: «número de balas gastadas»:

El conjunto de posibles valores que puede tomar la variable aleatoria es $\{1, 2, 3, \ldots\}$.

Este es un conjunto infinito numerable, pues no se conoce un máximo para el número de balas empleadas –que pudiera ser extremadamente grande para una persona con muy mala puntería–; es decir, X es una variable aleatoria discreta definida sobre un conjunto infinito numerable.

Más adelante se demostrará que también se cumple que $\sum_{i=1}^{\infty} p_i = 1$ con $p_i = \Pr\{X = i\}$.

Definición (de función de distribución) Sea X una variable aleatoria discreta, la función real F tal que

$$\forall t \in \mathbf{R}, \quad F(t) = \Pr(X \leq t)$$

se denomina función de distribución de la variable aleatoria X.

Propiedades

1. La función F es creciente, con $\lim_{t \to -\infty} F(t) = 0$ y $\lim_{t \to +\infty} F(t) = 1$.
2. $F(t) = \Pr(X \leq t) = \sum_{j \leq t} p_j$.
3. $\Pr(a < X \leq b) = F(b) - F(a)$.
4. $\Pr(a \leq X \leq b) = F(b) - F(a) + \Pr(X = a)$.

Observación. La probabilidad $\Pr(X = a)$ se calcula mediante $\Pr(X = a) = F(a) - F(a_-)$, donde $F(a_-)$ es el límite, por la izquierda, de la función de distribución en el punto a. Este concepto tiene importancia para el cálculo de las probabilidades en los puntos donde F tiene saltos.

Ejemplos

1. Continuando con el ejemplo del lanzamiento de tres monedas, se tiene que

$$\begin{aligned} F(0) &= p_0 = \frac{1}{8}, \\ F(1) &= p_0 + p_1 = \frac{1}{8} + \frac{3}{8} = \frac{1}{2}, \\ F(2) &= p_0 + p_1 + p_2 = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} = \frac{7}{8}, \\ F(3) &= p_0 + p_1 + p_2 + p_3 = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8} = 1. \end{aligned}$$

Con esto, la función de distribución es

$$F(t) = \begin{cases} 0, & \text{si } t < 0; \\ 1/8, & \text{si } 0 \leq t < 1; \\ 1/2, & \text{si } 1 \leq t < 2; \\ 7/8, & \text{si } 2 \leq t < 3; \\ 1, & \text{si } t \geq 3. \end{cases}$$

Los gráficos de las funciones de probabilidad y de distribución se dan en la Figura 3.2:

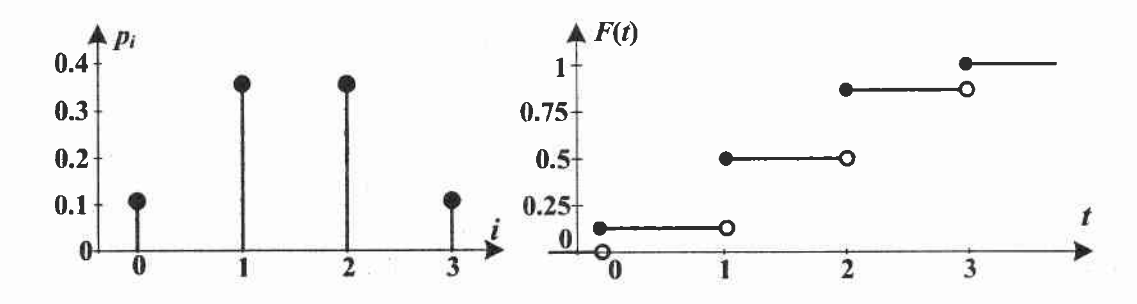


Figura 3.2:

2. En una prueba de calidad de un producto se tiene que en un lote de 12 piezas hay 8 buenas y 4 defectuosas. En el departamento de control de calidad se toma una muestra de 3 piezas. Construir la ley de la variable aleatoria «número de piezas buenas».

Solución: La variable aleatoria en cuestión puede tomar los valores 0, 1, 2 y 3; determinemos sus probabilidades.

El número de subconjuntos de 3 elementos obtenibles de un conjunto de 12 elementos es C_{12}^3, que es Card(Ω).

- Si $X = 0$, entonces todas las piezas son defectuosas, hay C_4^3 formas de escogerlas.
- Si $X = 1$, entonces 1 es buena y 2 son defectuosas, existen C_8^1 formas de escoger las piezas buenas y C_4^2 de escoger las defectuosas, entonces hay $\mathrm{C}_8^1\mathrm{C}_4^2$ formas de combinar las buenas y las defectuosas.
- Si $X = 2$, hay C_8^2 conjuntos de las piezas buenas y C_4^1 de defectuosas, para un total de $\mathrm{C}_8^2\mathrm{C}_4^1$ formas de combinarlas.
- Si $X = 3$, hay C_8^3 combinaciones de piezas buenas.

Entonces,

$$\Pr(X=0) = \frac{\mathrm{C}_4^3}{\mathrm{C}_{12}^3} = \frac{1}{55}, \qquad \Pr(X=1) = \frac{\mathrm{C}_8^1\mathrm{C}_4^2}{\mathrm{C}_{12}^3} = \frac{12}{55},$$
$$\Pr(X=2) = \frac{\mathrm{C}_8^2\mathrm{C}_4^1}{\mathrm{C}_{12}^3} = \frac{28}{55}, \qquad \Pr(X=3) = \frac{\mathrm{C}_8^3}{\mathrm{C}_{12}^3} = \frac{14}{55}.$$

Lo que se resume en su distribución de probabilidad:

k	0	1	2	3
$\Pr(X = k)$	1/55	12/55	28/55	14/55

Para definir una variable aleatoria no es necesario exhibir un fenómeno aleatorio particular, es suficiente dar una función de probabilidad o de distribución que cumpla las propiedades enunciadas.

3. La función de distribución de una variable aleatoria Y se define mediante:

$$F(t) = \begin{cases} 0, & \text{si } t < -3; \\ 1/2, & \text{si } -3 \leq t < 0; \\ 3/4, & \text{si } 0 \leq t < 2; \\ 1, & \text{si } t \geq 2. \end{cases}$$

Construir la tabla de distribución de probabilidad de Y.

Solución: De acuerdo a la definición de la función de distribución podemos ver que la variable aleatoria toma los valores -3, 0 y 2.

$$\begin{aligned} f(-3) &= \Pr(Y=-3) = F(-3) - F(-3_-) = \frac{1}{2} - 0 = \frac{1}{2}, \\ f(0) &= \Pr(Y=0) = F(0) - F(0_-) = \frac{3}{4} - \frac{1}{2} = \frac{1}{4}, \\ f(2) &= \Pr(Y=2) = F(2) - F(2_-) = 1 - \frac{3}{4} = \frac{1}{4}. \end{aligned}$$

De manera que la tabla de distribución de probabilidad es

k	-3	0	2
$\Pr(Y = k)$	1/2	1/4	1/4

4. Una variable aleatoria X está definida mediante la siguiente ley de probabilidad

x_k	1	2	3	4	5
$\Pr(X = x_k)$	0.3	0.2	0.1	0.15	0.25

a) Determinar $F(x)$.

b) Graficar $f(x)$ y $F(x)$.

c) Hallar: $\Pr(X = 1)$, $\Pr(X \leq 1)$, $\Pr(X < 1)$, $\Pr(1 < X \leq 2)$, $\Pr(1 \leq X \leq 2)$, $\Pr(1 < X < 2)$.

Solución:

a) Evidentemente, $F(x) = 0$ para $x < 1$, ahora

- Si $1 \leq x < 2$, $F(x) = p_1 = 0.3$.
- Si $2 \leq x < 3$, $F(x) = p_1 + p_2 = 0.5$.
- Si $3 \leq x < 4$, $F(x) = p_1 + p_2 + p_3 = 0.6$.
- Si $4 \leq x < 5$, $F(x) = p_1 + p_2 + p_3 + p_4 = 0.75$.
- Si $x \geq 5$, $F(x) = p_1 + p_2 + p_3 + p_4 + p_5 = 1$.

Es decir,

$$F(x) = \begin{cases} 0, & \text{si } x < 1; \\ 0.3, & \text{si } 1 \leq x < 2; \\ 0.5, & \text{si } 2 \leq x < 3; \\ 0.6, & \text{si } 3 \leq x < 4; \\ 0.75, & \text{si } 4 \leq x < 5; \\ 1, & \text{si } x \geq 5. \end{cases}$$

b) Los gráficos son:

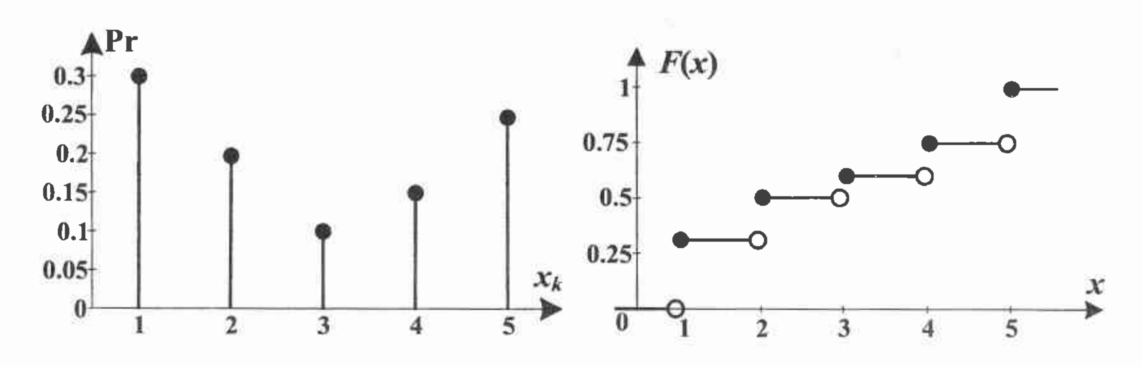


Figura 3.3:

c) Se tiene:

$$\begin{aligned} \Pr(X = 1) &= F(1) - F(1_-) = 0.3 - 0 = 0.3, \\ \Pr(X \leq 1) &= F(1) = 0.3, \\ \Pr(X < 1) &= F(1) - \Pr(X = 1) = 0.3 - 0.3 = 0, \\ \Pr(1 < X \leq 2) &= F(2) - F(1) = 0.5 - 0.3 = 0.2, \\ \Pr(1 \leq X \leq 2) &= F(2) - F(1) + \Pr(X = 1) = 0.5 - 0.3 + 0.3 = 0.5, \\ \Pr(1 < X < 2) &= F(2) - F(1) - \Pr(X = 2) = 0.2 - 0.2 = 0. \end{aligned}$$

3.1.2. Variables aleatorias continuas

Definición (de variable aleatoria continua) La variable aleatoria cuyo recorrido es un intervalo finito o infinito de **R** se llama variable aleatoria continua.

También, se dice que una variable aleatoria X es continua si para todo valor real x se tiene que

$$\Pr(X = x) = 0.$$

Usualmente, las variables continuas representan mediciones; por ejemplo, la estatura de una persona, el tiempo que se demora un programa en buscar un registro en una base de datos, la cantidad de sangre que tiene un animal.

Definición (de función de distribución) Sea X una variable aleatoria continua, la función real F tal que

$$\forall t \in \mathbf{R}, \quad F(t) = \Pr(X \leq t)$$

se denomina función de distribución de la variable aleatoria X.

Propiedades

1. F es creciente, con $\lim\limits_{t \to -\infty} F(t) = 0$ y $\lim\limits_{t \to +\infty} F(t) = 1$.

2. $\Pr(a \leq X \leq b) = \Pr(a < X \leq b) = \Pr(a \leq X < b) = \Pr(a < X < b) = F(b) - F(a)$.

Definición (de función de densidad) La función de densidad de una variable aleatoria continua X es una función real f que cumple:

a) $f(x) \geq 0$ para cualquier valor x.

b) $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

c) Para cualquier intervalo $A = [a, b]$, se tiene que

$$\Pr(A) = \Pr(a \leq X \leq b) = \int_a^b f(x)\,dx.$$

Veamos cómo están relacionadas las funciones de distribución y de densidad. (Ver Figura 3.4)

Teorema. Si F y f son las funciones de distribución y de densidad de la variable aleatoria X, respectivamente, ellas están ligadas mediante las igualdades

$$F(x) = \int_{-\infty}^{x} f(t)\,dt \qquad \text{y} \qquad f(x) = F'(x).$$

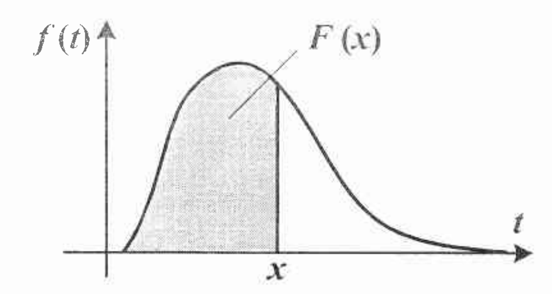


Figura 3.4: Relación entre las funciones de densidad y de distribución.

Tengamos presente que para el cálculo de probabilidades se emplea la siguiente equivalencia:

$$\Pr(a \leq X \leq b) = \int_a^b f(x)\,dx = F(b) - F(a).$$

En las variables aleatorias continuas es suficiente indicar la función de densidad o la función de distribución para que la variable aleatoria quede completamente definida.

Ejemplos

1. Una variable aleatoria X está definida mediante la función de distribución

$$F(x) = \begin{cases} 0, & \text{para } x \leq -1; \\ \frac{3}{4}x + \frac{3}{4}, & \text{para } -1 < x < \frac{1}{3}; \\ 1, & \text{para } x \geq \frac{1}{3}. \end{cases}$$

Hallar la probabilidad de que la variable aleatoria X tome un valor en el intervalo $\left(0, \frac{1}{3}\right)$.

Solución: La probabilidad de que X tome un valor en (a, b) es $\Pr(a < X < b) = F(b) - F(a)$. Si $a = 0$, $b = \frac{1}{3}$ obtenemos

$$\begin{aligned} \Pr\left(0 < X < \frac{1}{3}\right) &= F\left(\frac{1}{3}\right) - F(0) \\ &= \left[\frac{3}{4}x + \frac{3}{4}\right]_{x=1/3} - \left[\frac{3}{4}x + \frac{3}{4}\right]_{x=0} \\ &= \left[\frac{3}{4}\left(\frac{1}{3}\right) + \frac{3}{4}\right] - \left[\frac{3}{4}(0) + \frac{3}{4}\right] \\ &= \frac{1}{4}. \end{aligned}$$

2. La función de densidad de una variable aleatoria está dada por $f(x) = \alpha e^{-3x}$ en el intervalo $(0, \infty)$ y $f(x) = 0$ fuera de este intervalo. Hallar el valor de la constante α para que $f(x)$ así definida sea una función de densidad.

Solución: Primero verifiquemos que $f(x) \geq 0$.

- En $(-\infty, 0]$ no hay problema pues $f(x) = 0$.

- En $(0, \infty)$ se debe tener $f(x) \geq 0$, es decir $\alpha e^{-3x} \geq 0$. Pero $\forall x \in (0, \infty)$, $e^{-3x} \geq 0$; entonces, se debe tener que $\alpha \geq 0$.

Ahora, verifiquemos que $\displaystyle\int_{-\infty}^{\infty} f(x)dx = 1$, o sea $\displaystyle\int_0^{\infty} \alpha e^{-3x}dx = 1$.

La integral

$$\begin{aligned}\int_0^{\infty} \alpha e^{-3x}dx &= \alpha \int_0^{\infty} e^{-3x}dx = \alpha \left[\frac{-1}{3}e^{-3x}\right]_0^{\infty} \\ &= \frac{-\alpha}{3}[0-1] = \frac{\alpha}{3}.\end{aligned}$$

Consecuentemente, $\dfrac{\alpha}{3} = 1$; entonces $\alpha = 3$.

3. Dada la función de densidad de la variable aleatoria continua X:

$$f(x) = \begin{cases} 0, & \text{si } x \leq 0; \\ \cos x, & \text{si } 0 < x \leq \pi/2; \\ 0, & \text{si } x > \pi/2. \end{cases}$$

a) Hallar la función de distribución $F(x)$.

b) Determinar: $\Pr\left(\frac{\pi}{6} \leq X < \frac{\pi}{3}\right)$, $\Pr\left(X > \frac{\pi}{4}\right)$, $\Pr\left(\frac{\pi}{3} \leq X \leq \frac{11\pi}{12}\right)$.

Solución:

a) Utilizando la fórmula $F(x) = \displaystyle\int_{-\infty}^{x} f(t)\,dt$:

- Si $x \leq 0$, $f(x) = 0$, de manera que

$$F(x) = \int_{-\infty}^{x} 0\,dt = 0.$$

- Si $0 < x \leq \pi/2$, $f(x) = \cos x$, entonces

$$F(x) = \int_{-\infty}^{0} 0\,dt + \int_0^{x} \cos t\,dt = \operatorname{sen} x.$$

- Si $x > \pi/2$, $f(x) = 0$,

$$F(x) = \int_{-\infty}^{0} 0\,dt + \int_0^{\pi/2} \cos t\,dt + \int_{\pi/2}^{x} 0\,dt = \operatorname{sen} x\Big|_0^{\pi/2} = 1.$$

La función de distribución es:

$$F(x) = \begin{cases} 0, & \text{si } x \leq 0; \\ \operatorname{sen} x, & \text{si } 0 < x \leq \dfrac{\pi}{2}; \\ 1, & \text{si } x > \dfrac{\pi}{2}. \end{cases}$$

b) Para calcular las probabilidades emplearemos la función de distribución.

- $\Pr(a \leq X < b) = F(b) - F(a)$. Si $a = \dfrac{\pi}{6}$ y $b = \dfrac{\pi}{3}$:

$$\begin{aligned}\Pr\left(\frac{\pi}{6} \leq X < \frac{\pi}{3}\right) &= F\left(\frac{\pi}{3}\right) - F\left(\frac{\pi}{6}\right) \\ &= \operatorname{sen}\left(\frac{\pi}{3}\right) - \operatorname{sen}\left(\frac{\pi}{6}\right) = \frac{\sqrt{3}}{2} - \frac{1}{2} \\ &= 0.36603.\end{aligned}$$

- $\Pr(X > a) = 1 - \Pr(X \leq a) = 1 - F(a)$. Si $a = \dfrac{\pi}{4}$:

$$\begin{aligned}\Pr\left(X > \frac{\pi}{4}\right) &= 1 - F\left(\frac{\pi}{4}\right)\\ &= 1 - \operatorname{sen}\left(\frac{\pi}{4}\right) = 1 - \frac{\sqrt{2}}{2}\\ &= 0.29289.\end{aligned}$$

- $\Pr(a \leq X \leq b) = F(b) - F(a)$, con $a = \dfrac{\pi}{3}$ y $b = \dfrac{11\pi}{12}$:

$$\begin{aligned}\Pr\left(\frac{\pi}{3} \leq X \leq \frac{11\pi}{12}\right) &= F\left(\frac{11\pi}{12}\right) - F\left(\frac{\pi}{3}\right)\\ &= 1 - \operatorname{sen}\left(\frac{\pi}{3}\right) = 1 - \frac{\sqrt{3}}{2}\\ &= 0.13397.\end{aligned}$$

4. La función de densidad de una variable aleatoria T está dada por $f(t) = t - \dfrac{1}{2}$ en $(1, 2)$ y $f(t) = 0$ fuera de este intervalo. Hallar las probabilidades:

a) $\Pr(0 < T < 1.8)$;
b) $\Pr(1.2 < T \leq 1.5)$;
c) $\Pr(T > 1.5)$;
d) $\Pr(1.4 < T < 3)$.

Solución: Para el cálculo de las probabilidades utilizaremos la función de densidad.

a) $\Pr(0 < T < 1.8)$:

$$\begin{aligned}\Pr(0 < T < 1.8) &= \int_0^{1.8} f(t)\,dt = \int_0^1 f(t)\,dt + \int_1^{1.8} f(t)\,dt\\ &= \int_0^1 0\,dt + \int_1^{1.8}\left(t - \frac{1}{2}\right)dt = 0 + \frac{1}{2}\left(t^2 - t\right)\Big|_1^{1.8}\\ &= \frac{1}{2}\left[((1.8)^2 - 1.8) - 0\right] = 0.72.\end{aligned}$$

b) $\Pr(1.2 < T \leq 1.5)$:

$$\begin{aligned}\Pr(1.2 < T \leq 1.5) &= \int_{1.2}^{1.5} f(t)\,dt\\ &= \int_{1.2}^{1.5}\left(t - \frac{1}{2}\right)dt = \frac{1}{2}\left(t^2 - t\right)\Big|_{1.2}^{1.5}\\ &= \frac{1}{2}\left[((1.5)^2 - 1.5) - ((1.2)^2 - 1.2)\right] = 0.255.\end{aligned}$$

c) $\Pr(T > 1.5)$:

$$\begin{aligned}\Pr(T > 1.5) &= 1 - \Pr(T < 1.5) = 1 - \left(\int_{-\infty}^{1} 0\,dt + \int_1^{1.5}\left(t - \frac{1}{2}\right)dt\right)\\ &= 1 - \frac{1}{2}\left(t^2 - t\right)\Big|_1^{1.5} = 1 - \frac{1}{2}\left[((1.5)^2 - 1.5) - (1^2 - 1)\right]\\ &= 1 - \frac{1}{2}(0.75) = 0.625.\end{aligned}$$

d) $\Pr(1.4 < T < 3)$:

$$\begin{aligned}\Pr(1.4 < T < 3) &= \int_{1.4}^{3} f(t)\,dt = \int_{1.4}^{2} f(t)\,dt + \int_{2}^{3} f(t)\,dt \\ &= \int_{1.4}^{2} \left(t - \frac{1}{2}\right) dt + \int_{2}^{3} 0\,dt = \frac{1}{2}\left(t^2 - t\right)\Big|_{1.4}^{2} + 0 \\ &= \frac{1}{2}\left[(2^2 - 2) - \left((1.4)^2 - 1.4\right)\right] = 0.72.\end{aligned}$$

5. Hallar la función de densidad $f(x)$ de una variable aleatoria cuya función de distribución es

$$F(x) = \begin{cases} 0, & \text{si } x \leq 0; \\ \operatorname{sen} 2x, & \text{si } 0 < x \leq \frac{\pi}{4}; \\ 1, & \text{si } x > \frac{\pi}{4}. \end{cases}$$

Solución: Utilicemos la relación $f(x) = F'(x)$.

- Si $x \leq 0$, $F(x) = 0$, entonces $f(x) = F'(x) = 0$.
- Si $0 < x \leq \frac{\pi}{4}$, $F(x) = \operatorname{sen} 2x$, entonces $f(x) = F'(x) = 2\cos 2x$.
- Si $x > \frac{\pi}{4}$, $F(x) = 1$, entonces $f(x) = F'(x) = 0$.

Es decir,

$$f(x) = \begin{cases} 2\cos 2x, & \text{si } 0 < x \leq \frac{\pi}{4}; \\ 0, & \text{caso contrario.} \end{cases}$$

6. La vida útil de un elemento electrónico está dada por la función de densidad

$$f(t) = \begin{cases} \frac{1}{2}e^{-t/2}, & \text{si } t > 0; \\ 0, & \text{en lo demás;} \end{cases}$$

donde t es el tiempo (en horas). Calcular la probabilidad de que un elemento dure más de tres horas, dado que ya ha estado en uso más de dos horas.

Solución: Nos interesa $\Pr(T > 3|T > 2)$, que según la fórmula de la probabilidad condicional se tiene:

$$\Pr(T > 3|T > 2) = \frac{\Pr(T > 3)}{\Pr(T > 2)}$$

porque la intersección de los eventos $(T > 3)$ y $(T > 2)$ es el evento $(T > 3)$. Entonces,

$$\frac{\Pr(T > 3)}{\Pr(T > 2)} = \frac{\displaystyle\int_3^{\infty} \frac{1}{2}e^{-t/2}}{\displaystyle\int_2^{\infty} \frac{1}{2}e^{-t/2}} = \frac{e^{-3/2}}{e^{-1}} = e^{-1/2} = 0.606.$$

3.2. Distribuciones de funciones de variables aleatorias

Sea g una función real cuyo dominio contiene el recorrido de la variable aleatoria X, podemos definir una nueva variable aleatoria Y mediante

$$Y = g(X),$$

lo que quiere decir que, si la variable aleatoria X está definida según

$$\begin{aligned} X : \Omega &\longrightarrow A \subseteq \mathbf{R} \\ \omega &\longmapsto X(\omega) \end{aligned}$$

y g por

$$\begin{aligned} g : B &\longrightarrow \mathbf{R} \quad \text{con } A \subseteq B \\ x &\longmapsto g(x) \end{aligned}$$

la variable aleatoria Y se define por

$$\begin{aligned} Y : \Omega &\longrightarrow \mathbf{R} \\ \omega &\longmapsto Y(\omega) = g(X(\omega)). \end{aligned}$$

Si conocemos la ley de distribución F_X de X, vamos a determinar la ley de distribución F_Y de Y.

- Si X es variable aleatoria discreta,

  $$\Pr(Y = y_i) = \Pr(X = x_i),$$

  donde $y_i = g(x_i)$.

- Si X es una variable aleatoria continua. Supongamos que g es una función continua y estrictamente creciente en todo el eje real; entonces, existe la función inversa de g que la llamaremos h. Ésta también es continua y estrictamente creciente, por lo que

  $$y = g(x) \quad \text{si, y solo si} \quad x = h(y).$$

  Examinemos la definición de F_Y:

  $$F_Y(t) = \Pr(Y \leq t) = \Pr(g(X) \leq t).$$

  Aplicando la función inversa a los dos miembros de la desigualdad del argumento de la última expresión se obtiene

  $$\Pr(g(X) \leq t) = \Pr(X \leq h(t)) = F_X(h(t)).$$

  Luego, se tiene la siguiente equivalencia entre las funciones de distribución de X y de Y:

  $$F_Y(t) = F_X(h(t)).$$

  Si las funciones F_X y h son derivables, se pueden derivar ambos miembros de la igualdad anterior, empleando la regla de la cadena:

  $$F_Y'(t) = F_X'(h(t)) \cdot h'(t).$$

  Ello nos conduce a la siguiente relación entre las funciones de densidad:

  $$f_Y(t) = f_X(h(t)) \cdot h'(t).$$

  Observación. Si la función g no es monótona en el intervalo de los posibles valores de X, hay que dividir este intervalo en subintervalos tales que g sea monótona y aplicar el resultado anterior.

Ejemplos

1. Dada la función de distribución F_X de la variable aleatoria X, hallar las funciones de distribución y de densidad de $Y = aX + b$, para: a) $a > 0$; b) $a < 0$.

   Solución:

   a) Sea $Y = aX + b$, $a > 0$. Tenemos:

   $$\begin{aligned} F_Y(t) &= \Pr(Y \leq t) = \Pr(aX + b \leq t) \\ &= \Pr\left(X \leq \frac{t-b}{a}\right) = F_X\left(\frac{t-b}{a}\right). \end{aligned}$$

   Además,

   $$g(t) = at + b, \qquad h(t) = \frac{t-b}{a}, \qquad h'(t) = \frac{1}{a},$$

   También,

   $$f_Y(t) = \frac{1}{a} f_X\left(\frac{t-b}{a}\right).$$

   b) Si $Y = aX + b$, con $a < 0$, resulta que $g(t) = at + b$, que es continua, pero estrictamente decreciente; sin embargo, podemos hallar la función de distribución de Y mediante su definición:

   $$\begin{aligned} F_Y(t) &= \Pr(Y \leq t) = \Pr(aX + b \leq t) \\ &= \Pr\left(X \geq \frac{t-b}{a}\right) = 1 - \Pr\left(X < \frac{t-b}{a}\right) \\ &= 1 - F_X\left(\frac{t-b}{a}\right) + \Pr\left(X = \frac{t-b}{a}\right). \end{aligned}$$

   La función de densidad es

   $$f_Y(t) = F_Y'(t) = -\frac{1}{a} f_X\left(\frac{t-b}{a}\right).$$

2. Considere la variable aleatoria $Y = X^2$. Hallar su función de distribución.

   Solución: Aquí, $g(x) = x^2$, función que no es estrictamente creciente. Por definición tenemos:

   $$F_Y(t) = \Pr(Y \leq t) = \Pr(X^2 \leq t).$$

   - Si $t < 0$, el evento «$X^2 \leq t$» es vacío; por tanto, $\Pr(X^2 \leq t) = 0$, lo que implica que $F_Y(t) = 0$.
   - Si $t \geq 0$,

   $$\begin{aligned} F_Y(t) &= \Pr(X^2 \leq t) = \Pr(|X| \leq \sqrt{t}) \\ &= \Pr(-\sqrt{t} \leq X \leq \sqrt{t}) \\ &= F_X(\sqrt{t}) - F_X(-\sqrt{t}) + \Pr(X = -\sqrt{t}). \end{aligned}$$

   Si F_X es continua se tiene la siguiente expresión para F_Y:

   $$F_Y(t) = \begin{cases} 0, & \text{si } t < 0; \\ F_X(\sqrt{t}) - F_X(-\sqrt{t}), & \text{si } t \geq 0. \end{cases}$$

   Si F_X es derivable en $\sqrt{t}$ y en $-\sqrt{t}$, se obtiene la función de densidad:

   $$f_Y(t) = \begin{cases} 0, & \text{si } t < 0; \\ \dfrac{f_X(\sqrt{t}) + f_X(-\sqrt{t})}{2\sqrt{t}}, & \text{si } t > 0. \end{cases}$$

3. Determinar la ley de distribución de la variable aleatoria $Y = X^2$, si X está definida mediante:

X	-3	-2	0	2
$\Pr(X = k)$	$1/6$	$1/4$	$1/4$	$1/3$

Solución: Como la variable aleatoria es discreta, basta aplicar la relación $\Pr(Y = y_i) = \Pr(X = x_i)$, a cada uno de los valores que toma $Y = X^2$. Entonces, tenemos que

X^2	$(-3)^2$	$(-2)^2$	0^2	2^2
$\Pr\left(X^2 = k\right)$	$1/6$	$1/4$	$1/4$	$1/3$

Es decir,

Y	9	4	0	4
$\Pr(Y = k)$	$1/6$	$1/4$	$1/4$	$1/3$

Como el valor $Y = 4$ se repite 2 veces, unificamos sus probabilidades y la tabla queda así:

Y	0	4	9
$\Pr(Y = k)$	$1/4$	$7/12$	$1/6$

4. Hallar la función de distribución de la variable aleatoria $Y = e^{-X}$, si X está definida mediante

$$F_X(t) = \begin{cases} 0, & \text{si } t < -1; \\ \dfrac{t+1}{2}, & \text{si } -1 \leq t \leq 1; \\ 1, & \text{si } t > 1. \end{cases}$$

Solución: Se tienen las siguientes igualdades:

$$\begin{aligned} F_Y(t) &= \Pr(Y \leq t) = \Pr\left(e^{-X} \leq t\right) \\ &= \Pr(X \geq -\ln t) = 1 - \Pr(X < -\ln t) \\ &= 1 - F_X(-\ln t). \end{aligned}$$

Por otro lado,

$$\begin{aligned} F_X(-\ln t) &= \begin{cases} 0, & \text{si } -\ln t < -1; \\ \dfrac{-\ln t + 1}{2}, & \text{si } -1 \leq -\ln t \leq 1; \\ 1, & \text{si } -\ln t > 1. \end{cases} \\ &= \begin{cases} 0, & \text{si } t > e; \\ \dfrac{-\ln t + 1}{2}, & \text{si } e^{-1} \leq t \leq e; \\ 1, & \text{si } t < e^{-1}. \end{cases} \end{aligned}$$

Por tanto,

$$F_Y(t) = 1 - F_X(-\ln t) = \begin{cases} 1, & \text{si } t > e; \\ \dfrac{1 + \ln t}{2}, & \text{si } e^{-1} \leq t \leq e; \\ 0, & \text{si } t < e^{-1}. \end{cases}$$

3.3. Ejercicios

1. Indique si las siguientes variables aleatorias son discretas o continuas y su rango de definición:

   a) El número de bytes defectuosos en el disco duro de una computadora de 100 Gb;

   b) La distancia de lanzamiento de la jabalina por un atleta;

   c) El número de goles que anota un equipo de fútbol en un partido;

   d) La cantidad de dinero, en dólares, ganada (o perdida) por un apostador;

   e) El tiempo de uso diario de una computadora;

   f) El tiempo de espera del autobús en una parada;

   g) El número de años que sobrevive una persona a la muerte de su cónyuge;

   h) La variación en el tiempo de sueño de una persona sometida a un tratamiento.

2. Indique al menos tres variables aleatorias discretas y tres variables aleatorias continuas. Especifique su rango de definición.

3. Se arroja un dado y se designan por $A = \{$el número de los puntos aparecidos es par$\}$ y por $B = \{$el número de los puntos aparecidos se divide por 3$\}$. Para los dos eventos, halle la ley de distribución y grafíquelas.

4. Determine la función de distribución de la variable aleatoria X que está definida por la ley que se presenta en la tabla.

X	-2	0	$\sqrt{3}$
p	$1/4$	$2/3$	$1/12$

5. Un escritor ha lanzado al mercado una nueva novela. La probabilidad de que la novela sea muy exitosa es 0.6, de que sea medianamente exitosa es 0.3 y de que sea un fracaso es 0.1. Los beneficios esperados son: si la novela es muy exitosa, 100 mil dólares; si la novela es moderadamente existosa, 50 mil dólares; y, si es un fracaso, 10 mil dólares. Forme la ley de distribución de los beneficios esperados por el escritor.

6. Una agencia automotriz recibe un embarque de 20 automóviles nuevos; entre éstos, 2 tienen defectos. La agencia debe seleccionar, aleatoriamente, 3 automóviles de entre los 20 para venderlos. Forme la ley de distribución de la variable aleatoria «número de carros defectuosos entre los escogidos».

7. Un apuesto príncipe visita a un rey que tiene cuatro hijas casaderas, con la intención de integrarse en la familia. Las probabilidades que tiene el príncipe de ser aceptado por cada una de las princesas son 0.6, 0.8, 0.2 y 0.4. El príncipe pide la mano de cada una de ellas de forma consecutiva y se casa con la primera que acepte. Sea X la variable aleatoria definida como $X = i$ si se casa con la $i-$ésima hija ($i = 1, \ldots, 4$) y $X = 0$ si todas le rechazan. Calcule la ley de probabilidad de X y su función de distribución.

8. Una chapa para puertas consta de tres piezas mecánicas. Suponga que las probabilidades de que la primera, la segunda y la tercera piezas cumplan con las especificaciones son 0.95, 0.98 y 0.99, respectivamente. Determine la distribución de probabilidad del número de piezas que cumplen las especificaciones en una chapa.

9. Sea X una variable aleatoria discreta cuya función de probabilidad es $p(x) = \dfrac{k}{x^2}$, $x = 1, 2, 3, 4, 5$.

   a) Encuentre el valor de k para que la función $p(x)$ sea la función de probabilidad de X;

   b) Calcule $\Pr(1 < X \leq 4)$.

10. La función de probabilidad f de una variable aleatoria X es nula salvo en los puntos $t = 0$, 1 y 2. En ellos toma los valores:

$$f(0) = 4c^2, \quad f(1) = 4c - 10c^2, \quad f(2) = 4c - 1,$$

para un cierto valor de c.

a) Determine el valor de c;
b) Calcule: $\Pr(X < 1)$, $\Pr(X < 2)$, $\Pr(0 < X < 3)$.

11. Una variable aleatoria X se dice que sigue la ley de Benford si se cumple que

$$\Pr(X = k) = \log_{10}\left(1 + \frac{1}{k}\right), \quad k = 1, 2, \ldots, 9.$$

a) Verifique que es una función de probabilidad;
b) Calcule la probabilidad de obtener número impares;
c) Grafique la función de probabilidad.

12. Una variable aleatoria Y se define para un entero positivo fijo a ($a \geq 1$) cualquiera mediante

$$\Pr(Y = k) = \frac{a}{k-1} - \frac{a}{k}, \quad k = a+1, a+2, \ldots$$

a) Verifique que es una función de probabilidad;
b) Demuestre que $\Pr(Y > k) = \dfrac{a}{k}$ para $k = a, a+1, \ldots$;
c) Fije un valor para el parámetro a y grafique la función de probabilidad.

13. Una variable aleatoria discreta X está definida según la ley

$$\Pr(X = k) = p(1-p)^k, \quad k = 0, 1, 2, \ldots \text{ y } p \in (0, 1).$$

a) Verifique que es una función de probabilidad;
b) Determine la función de distribución;
c) Calcule: $\Pr(X > 2)$, $\Pr(X \geq 4)$, $\Pr(X < 3)$.

14. Dadas las funciones de densidad f, encuentre el valor de la constante c de tal manera que ellas estén bien definidas.

a) $f(x) = \begin{cases} cx(x-1), & \text{si } 2 \leq x < 4; \\ 0, & \text{caso contrario.} \end{cases}$

b) $f(x) = 1 - c|1 - x|$, si $0 < x < 2$;

c) $f(x) = \begin{cases} ce^{-x}, & \text{si } x \geq 0; \\ ce^{x}, & \text{si } x < 0. \end{cases}$

d) $f(x) = cxe^{-x/2}$, si $0 < x < \infty$;

e) $f(x) = \dfrac{c}{1+x^2}$, $-\infty < x < \infty$;

f) $f(x) = \begin{cases} c \operatorname{sen} x, & \text{si } 0 \leq x \leq \frac{\pi}{2}; \\ 0, & \text{caso contrario.} \end{cases}$

15. Dada la función de distribución de una variable aleatoria X:

$$F(x) = \begin{cases} 0, & \text{si } x < 0; \\ 1/4, & \text{si } 0 \leq x < 1; \\ 1/3, & \text{si } 1 \leq x < 2; \\ x/6, & \text{si } 2 \leq x < 4; \\ (x-2)/3, & \text{si } 4 \leq x < 5; \\ 1, & \text{si } x \geq 5. \end{cases}$$

Calcule las probabilidades:

a) $\Pr(1 \leq X \leq 5)$;

b) $\Pr(2 < X \leq 4)$;

c) $\Pr(0 < X < 3)$;

d) $\Pr(4 \leq X < 6)$.

16. Se tiene la función de distribución de una variable aleatoria definida por

$$F(x) = \begin{cases} 0, & \text{si } x < -\sqrt{2}; \\ 1/8, & \text{si } -\sqrt{2} \leq x < 0; \\ 2/5, & \text{si } 0 \leq x < 1; \\ 1/2, & \text{si } 1 \leq x < \sqrt{2}; \\ 3/4, & \text{si } \sqrt{2} \leq x < 5/2; \\ 1, & \text{si } x \geq 5/2. \end{cases}$$

Determine la función de probabilidad asociada y grafíquela.

17. La función de densidad de una variable aleatoria X está definida mediante

$$f(x) = \begin{cases} 0, & \text{si } x \leq \dfrac{\pi}{6}; \\ 3 \operatorname{sen} 3x, & \text{si } \dfrac{\pi}{6} < x \leq \dfrac{\pi}{3}; \\ 0, & \text{si } x > \dfrac{\pi}{3}. \end{cases}$$

a) Halle la función de distribución F;

b) Determine: $\Pr(X = 0.2)$, $\Pr(X \leq \pi/4)$, $\Pr(X > \pi/3)$, $\Pr(\pi/12 \leq X \leq \pi)$.

18. Una variable aleatoria X tiene distribución continua F, siendo

$$F(t) = \begin{cases} 0, & \text{si } t \leq 0; \\ ct, & \text{si } 0 < t \leq 1; \\ 1, & \text{si } t > 1. \end{cases}$$

a) Determine la constante c y halle la función de densidad f;

b) Calcule las probabilidades $\Pr(X = 1/3)$, $\Pr(X \leq 1/3)$, $\Pr(|X| < 1/4)$.

19. Considere una variable aleatoria continua Z con densidad de probabilidad

$$f(z) = \begin{cases} (1+b)z^b, & \text{si } z \in [0, a]; \\ 0, & \text{si } z \notin [0, a]. \end{cases}$$

a) Calcule los valores de los parámetros a y b sabiendo que $\Pr\left(Z \leq \dfrac{1}{2}\right) = \dfrac{1}{8}$;

b) Encuentre la función de distribución de Z.

20. Una variable aleatoria X tiene por función de distribución a

$$F(x) = \begin{cases} 0, & \text{si } x < -2; \\ ax + b, & \text{si } -2 \leq x \leq 2; \\ 1, & \text{si } x > 2. \end{cases}$$

a) Determine los valores de a y b;

b) Encuentre la densidad f;

c) Halle: $\Pr(X < 0)$, $\Pr(|X| < 1.5)$, $\Pr(|X| > 1.2)$.

21. El tiempo en minutos que una persona espera un autobús es una variable aleatoria cuya función de densidad viene dada por las fórmulas: $f(t) = \dfrac{1}{2}$ para $0 \leq t < 1$, $f(t) = \dfrac{1}{6}$ para $1 \leq t < 4$, $f(t) = 0$ para los demás valores de t. Calcule la probabilidad de que el tiempo de espera sea:

a) mayor que un minuto;

b) menor que dos minutos;

c) mayor que tres minutos.

22. Los registros de ventas diarias de una empresa que comercializa computadoras muestran que venderán 0, 1 o 2 computadoras de acuerdo a la siguiente tabla:

No. de ventas	0	1	2
Probabilidad	0.7	0.2	0.1

a) Determine la distribución de probabilidad de X, el número de ventas;

b) Calcule la probabilidad de que al menos se realice una venta en el día.

23. Un blanco circular de radio 1 se divide en 5 anillos circulares por medio de 5 discos concéntricos de radios: $\frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}$ y 1. Un jugador lanza un dardo al blanco, si el dardo alcanza el anillo circular comprendido entre los círculos de radios $\frac{k}{5}$ y $\frac{k+1}{5}$, $(k = 0, 1, 2, 3, 4)$, tiene k puntos y gana $5 - k$ dólares. Determine las distribuciones de probabilidad:

a) del puntaje del jugador;

b) de la ganancia del jugador.

24. Una empresa alquila el tiempo de cómputo de un tipo especial de computadora a una universidad. La empresa debe planear su presupuesto, por lo que ha estudiado el tiempo de empleo de la computadora. El tiempo semanal de alquiler (en horas) sigue la función de densidad dada por:

$$f(t) = \begin{cases} \frac{3}{64}t^2(4-t), & \text{si } 0 \le t \le 4; \\ 0, & \text{caso contrario.} \end{cases}$$

a) Determine la función de distribución del tiempo de empleo de la computadora;

b) Calcule la probabilidad de que el tiempo de uso de la computadora, en una semana, sea mayor que 2 horas;

c) El presupuesto de la empresa solo cubre 3 horas de tiempo semanal de uso de la computadora. ¿Con qué frecuencia se rebasará ese límite de presupuesto?;

d) ¿Cuánto tiempo de alquiler se debe presuponer por semana si esta cifra solo se puede rebasar con una probabilidad de 0.1?

25. La cantidad de pan (en cientos de kilogramos) que vende una panadería en un día es una variable aleatoria con función de densidad

$$f(x) = \begin{cases} cx, & \text{si } 0 \le x < 3; \\ c(6-x), & \text{si } 3 \le x < 6; \\ 0, & \text{caso contrario.} \end{cases}$$

a) Encuentre el valor de c;

b) ¿Cuál es la probabilidad que el número de kilos de pan que se vende en un día sea: (i) más de 300 kg?, (ii) entre 150 y 450 kg?;

c) Denote por A y B los eventos definidos en (i) e (ii), respectivamente. ¿Son independientes A y B?

26. La cantidad (en gramos) de fertilizante químico que una planta puede recibir es una variable aleatoria cuya función de densidad es

$$f(x) = \begin{cases} \frac{3x(8-x)}{256}, & \text{si } x \in [0, 8]; \\ 0, & \text{caso contrario.} \end{cases}$$

a) Halle la probabilidad de que la planta reciba menos de 3 gramos;

b) Si la planta muere si recibe más de 6 g, ¿cuál es la probabilidad de que la planta muera por exceso de fertilizante?;

c) Si se trata de establecer una norma para la cantidad de fertilizante utilizada, ¿cuál es la cantidad máxima recomendada utilizar para que solo se sobrepase esta cantidad el 35 % de las veces?

27. Se extrae una bolita al azar de un bolillero que contiene 3 bolitas numeradas de 1 a 3. Llamamos X al número de la bolita extraída. Una vez conocido el valor de X, extraemos una nueva bolita al azar de otro bolillero que contiene $4 - X$ bolitas numeradas de X a 3 (por ejemplo: si $X = 2$, la segunda bolita se extrae de un bolillero que contiene dos bolitas con los números 2 y 3). Llamamos Y al número de la bolita extraída en el segundo bolillero.

a) Calcule $\Pr(Y = 3|X = 1)$;

b) Calcule $\Pr(Y = 3)$;

c) ¿Son X y Y independientes? Justifique;

d) Halle la distribución de probabilidad de Y.

28. Una variable aleatoria X tiene densidad

$$f(x) = \begin{cases} 1, & \text{si } x \in [0, 1]; \\ 0, & \text{si } x \notin [0, 1]. \end{cases}$$

a) Si $Y = X^2$, halle la función de distribución de Y;

b) Calcule las probabilidades: $\Pr\left(\frac{1}{16} < X^2 < \frac{1}{9}\right)$ y $\Pr\left(\frac{1}{8} < Y < \frac{5}{6}\right)$.

29. Una variable aleatoria Z tiene función de densidad

$$f(z) = \begin{cases} \frac{1}{2}, & \text{si } z \in [-1, 1]; \\ 0, & \text{si } z \notin [-1, 1]. \end{cases}$$

Halle la ley de la variable $T = -5Z$.

30. Una variable aleatoria X tiene función de densidad

$$f(x) = \begin{cases} \frac{1}{4}, & \text{si } x \in [-2, 2]; \\ 0, & \text{si } x \notin [-2, 2]. \end{cases}$$

Halle la probabilidad $\Pr(X^2 < 1)$.

31. Una variable aleatoria Y está distribuida según la ley

$$f(y) = \begin{cases} \frac{1}{3}, & \text{si } y \in [-1, 2]; \\ 0, & \text{caso contrario.} \end{cases}$$

Halle la función de densidad de la variable $U = Y^2$.

32. Una variable aleatoria X tiene densidad $f_X(x) = e^{-x}$, si $x \geq 0$. Encuentre las funciones de distribución y de densidad de la variable aleatoria $Z = e^{-X}$.

33. **Una variable aleatoria X tiene función de distribución $F_X(x) = 1 - e^{-\alpha x}$, si $x \geq 0$. Halle las funciones de densidad de:**

a) $Y = \sqrt{X}$; b) $Z = \dfrac{1}{\alpha} \ln X$.

En las secciones precedentes vimos que una variable aleatoria queda definida por su función de distribución, pero muchas veces solo se desea tener una idea del comportamiento general de las variables aleatorias, sin dar detalles de su distribución de probabilidad; para tal propósito, examinaremos dos características teóricas de las variables aleatorias: la esperanza y la varianza, que son dos parámetros que miden la localización y la dispersión de los valores que toma la variable.

3.4. La esperanza matemática

La esperanza matemática –o simplemente *esperanza*– de una variable aleatoria X, se simboliza por $\mathbf{E}(X)$ y su definición es la siguiente:

Definición (de esperanza de una variable aleatoria discreta) Sea X una variable aleatoria discreta, la esperanza es un número real que se calcula según:

1. Si X toma un número finito de valores $x_1, x_2, \ldots, x_n$ con probabilidades $p_1 = \Pr(X = x_1)$, $p_2 = \Pr(X = x_2), \ldots, p_n = \Pr(X = x_n)$:

$$\mathbf{E}(X) = \sum_{k=1}^{n} p_k \, x_k.$$

2. Si X toma un número infinito de valores $x_1, x_2, \ldots$ con probabilidades $p_k = \Pr(X = x_k)$, $k = 1, 2, \ldots$:

$$\mathbf{E}(X) = \sum_{k=1}^{\infty} p_k \, x_k.$$

Definición (de esperanza de una variable aleatoria continua) Sea X una variable aleatoria continua, cuya función de densidad es $f(x)$, la esperanza es un número real que se calcula según:

$$\mathbf{E}(X) = \int_{-\infty}^{\infty} x f(x) \, dx.$$

A la esperanza también se la denomina ***media poblacional*** o ***valor esperado*** de la variable aleatoria y se la suele notar como μ.

Observación. Si $f(x)$ toma valores distintos de cero en un intervalo $[a, b]$, la esperanza se calcula como

$$\mathbf{E}(X) = \int_{a}^{b} x f(x) \, dx.$$

La esperanza posee varias propiedades, independientes del tipo de la variable aleatoria. A continuación vamos a enunciarlas y demostrar algunas de ellas, en el caso de una variable aleatoria continua, los otros dos casos quedan como ejercicio para el lector.

Propiedades

1. La esperanza de una constante es el valor de la constante:

$$\mathbf{E}(c) = c, \quad c \text{ constante}.$$

Demostración:

$$\mathbf{E}(c) = \int_{-\infty}^{\infty} c f(x)\, dx = c \int_{-\infty}^{\infty} f(x)\, dx = c \cdot 1 = c.$$

2. **Aditividad**. La esperanza de la suma de dos variables aleatorias es igual a la suma de las esperanzas de los dos sumandos:

$$\mathbf{E}(X + Y) = \mathbf{E}(X) + \mathbf{E}(Y).$$

3. Un factor constante c se puede sacar del símbolo de la esperanza matemática:

$$\mathbf{E}(cX) = c\mathbf{E}(X).$$

Demostración:

$$\mathbf{E}(cX) = \int_{-\infty}^{\infty} cx f(x)\, dx = c \int_{-\infty}^{\infty} x f(x)\, dx = c\mathbf{E}(X).$$

4. Sea g una función real, la esperanza de la variable aleatoria $Y = g(X)$ está definida por

$$\mathbf{E}(Y) = \mathbf{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f(x)\, dx.$$

En particular si $g(x) = x^2$ se tiene

$$\mathbf{E}\left(X^2\right) = \int_{-\infty}^{\infty} x^2 f(x)\, dx.$$

5. Si X y Y son dos variables aleatorias independientes

$$\mathbf{E}(XY) = \mathbf{E}(X)\,\mathbf{E}(Y).$$

Observaciones:

1. Por las propiedades 2. y 3., si $Y = aX + b$, entonces

$$\mathbf{E}(Y) = a\mathbf{E}(X) + b.$$

2. Si la función de densidad es simétrica respecto a la recta $x = m$, entonces $\mathbf{E}(X) = m$. (Figura 3.5)

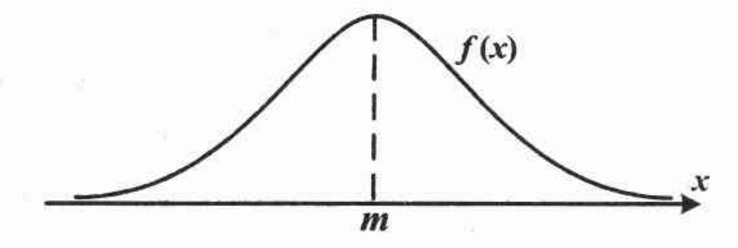


Figura 3.5: Función de densidad simétrica respecto a la recta $x = m$.

Dos variables aleatorias con la misma esperanza pueden tener distribuciones diferentes. Para diferenciarlas es necesario introducir otra característica teórica que informe sobre la dispersión de su posibles valores.

3.5. La varianza

La idea de esperanza no indica cómo está distribuida la masa en torno a su centro; ésto se expresa mediante la *varianza* de la variable aleatoria X, que se nota $\mathrm{Var}(X)$ o σ^2.

Definición (de varianza) La varianza de una variable aleatoria X es un número no negativo que se calcula por:

$$\mathrm{Var}(X) = \mathbf{E}(X - \mathbf{E}(X))^2,$$

o, equivalentemente, por

$$\mathrm{Var}(X) = \mathbf{E}(X^2) - (\mathbf{E}(X))^2.$$

Según el tipo de variable aleatoria, se calcula de la siguiente manera:

1. Para una variable aleatoria discreta que toma un número finito de valores x_1, x_2, ..., x_n con probabilidades $p_1 = \Pr(X = x_1)$, $p_2 = \Pr(X = x_2)$, ..., $p_n = \Pr(X = x_n)$:

$$\mathrm{Var}(X) = \sum_{k=1}^{n} [x_k - \mathbf{E}(X)]^2 p_k \quad \text{o} \quad \mathrm{Var}(X) = \sum_{k=1}^{n} p_k x_k^2 - (\mathbf{E}(X))^2.$$

2. Para una variable aleatoria discreta que toma un número infinito de valores x_1, x_2, ... con probabilidades $p_k = \Pr(X = x_k)$, $k = 1, 2, \ldots$:

$$\mathrm{Var}(X) = \sum_{k=1}^{\infty} [x_k - \mathbf{E}(X)]^2 p_k \quad \text{o} \quad \mathrm{Var}(X) = \sum_{k=1}^{\infty} p_k x_k^2 - (\mathbf{E}(X))^2.$$

3. Para una variable aleatoria continua con función de densidad $f(x)$:

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} [x - \mathbf{E}(X)]^2 f(x)\, dx \quad \text{o} \quad \mathrm{Var}(X) = \int_{-\infty}^{\infty} x^2 f(x)\, dx - (\mathbf{E}(X))^2.$$

Observación. Al igual que en la esperanza, si $f(x)$ está definida en $[a, b]$:

$$\mathrm{Var}(X) = \int_{a}^{b} [x - \mathbf{E}(X)]^2 f(x)\, dx.$$

La varianza da la idea de cuán ampliamente dispersos se encuentran, en torno de la media, los valores que toma la variable aleatoria:

1. Una mayor varianza indica que los valores tienden a estar más alejados de la media.
2. Una menor varianza indica que los valores tienden a estar más concentrados alrededor de la media.

Definición (de desviación estándar) La desviación estándar de una variable aleatoria X es igual a la raíz cuadrada de la varianza:

$$\sigma = \sqrt{\mathrm{Var}(X)}.$$

Propiedades

1. La varianza de una constante es cero. Es decir, para toda constante c:

$$\mathrm{Var}(c) = 0, \quad c \text{ constante.}$$

Demostración:

$$\begin{aligned}\mathrm{Var}(c) &= \int_{-\infty}^{\infty} [c - \mathbf{E}(c)]^2 f(x)\,dx = \int_{-\infty}^{\infty} [c - c]^2 f(x)\,dx \\ &= \int_{-\infty}^{\infty} 0\, f(x)\,dx = 0.\end{aligned}$$

2. Un factor constante c se puede sacar del símbolo de la varianza, elevándolo al cuadrado:

$$\mathrm{Var}(cX) = c^2\,\mathrm{Var}(X).$$

Demostración:

$$\begin{aligned}\mathrm{Var}(cX) &= \int_{-\infty}^{\infty} [cx - \mathbf{E}(cX)]^2 f(x)\,dx = \int_{-\infty}^{\infty} [cx - c\mathbf{E}(X)]^2 f(x)\,dx \\ &= \int_{-\infty}^{\infty} c^2[x - \mathbf{E}(X)] f(x)\,dx = c^2\,\mathrm{Var}(X).\end{aligned}$$

3. **Aditividad**. La varianza de la suma de dos variables aleatorias independientes es igual a la suma de las varianzas de los dos sumandos. Es decir, si X y Y son independientes

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

Demostración: En efecto,

$$\begin{aligned}\mathrm{Var}(X+Y) &= \mathbf{E}[(X+Y) - \mathbf{E}(X+Y)]^2 \\ &= \mathbf{E}[(X - \mathbf{E}(X)) + (Y - \mathbf{E}(Y))]^2 \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathbf{E}[(X - \mathbf{E}(X))(Y - \mathbf{E}(Y))].\end{aligned}$$

Como las variables X y Y son independientes, también lo son las cantidades $X - \mathbf{E}(X)$ y $Y - \mathbf{E}(Y)$, por lo tanto,

$$\mathbf{E}[(X - \mathbf{E}(X))(Y - \mathbf{E}(Y))] = \mathbf{E}[X - \mathbf{E}(X)] \cdot \mathbf{E}[Y - \mathbf{E}(Y)] = 0.$$

En consecuencia, $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Observación. De las propiedades 1. y 2. se verifica que

$$\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X).$$

Ejemplos

1. La variable aleatoria discreta X está definida según la ley

X	-4	6	10
p	0.2	0.3	0.5

Hallar la esperanza y la varianza de: a) la variable aleatoria X; b) la variable aleatoria $Y = 0.5X + 2$.

Solución:

a) $\mathbf{E}(X) = \sum_{k=1}^{3} x_k p_k = -4 \times 0.2 + 6 \times 0.3 + 10 \times 0.5 = 6.$

Calculemos $\mathbf{E}(X^2)$:

$$\mathbf{E}(X^2) = \sum_{k=1}^{3} x_k^2 p_k = (-4)^2(0.2) + (6)^2(0.3) + (10)^2(0.5) = 64.$$

Entonces,

$$\text{Var}(X) = \mathbf{E}(X^2) - (\mathbf{E}(X))^2 = 64 - (6)^2 = 28.$$

b) Vamos a aplicar las propiedades de la esperanza y la varianza para calcularlas:

$$\begin{aligned} \mathbf{E}(Y) &= \mathbf{E}(0.5X + 2) = 0.5\mathbf{E}(X) + \mathbf{E}(2) \\ &= 0.5 \times 6 + 2 = 5. \\ \text{Var}(Y) &= \text{Var}(0.5X + 2) = (0.5)^2 \,\text{Var}(X) \\ &= 0.25 \times 28 = 7. \end{aligned}$$

2. En una rifa se venden 300 boletos, a un dólar cada uno. El primer premio es 100 dólares, el segundo premio es 50 dólares y hay otros cinco premios de 10 dólares. ¿Cuál es la ganancia esperada de una persona que compra un boleto?

Solución: Sea X la cantidad ganada por un boleto premiado; entonces, X sigue la siguiente ley

X	100	50	10	0
p	1/300	1/300	5/300	293/300

Así,

$$\begin{aligned} \mathbf{E}(X) &= 100 \times \frac{1}{300} + 50 \times \frac{1}{300} + 10 \times \frac{5}{300} + 0 \times \frac{293}{300} \\ &= 0.67. \end{aligned}$$

Como la persona paga 1 dólar por el boleto, la ganancia (total) esperada es $\mathbf{E}(G) = 0.67 - 1 = -0.33$ dólares; es decir, una pérdida.

3. Una persona quiere abrir una puerta y tiene 5 llaves, de las cuales solo una corresponde a la cerradura. La persona va eligiendo al azar y probando abrir la puerta. Calcular la esperanza y la varianza del número de intentos si separa las llaves que probó anteriormente.

Solución: Como cada vez separa las llaves utilizadas, cada llave tiene la misma probabilidad de abrir la cerradura; por lo que la variable aleatoria X: «número de llaves utilizadas», sigue la siguiente ley:

X	1	2	3	4	5
p_i	1/5	1/5	1/5	1/5	1/5

Entonces,

$$\begin{aligned} \mathbf{E}(X) &= \sum_{k=1}^{5} k\, p_k = \frac{1}{5}\sum_{k=1}^{5} k = 3, \\ \mathbf{E}\left(X^2\right) &= \sum_{k=1}^{5} k^2 p_k = \frac{1}{5}\sum_{k=1}^{5} k^2 = 11, \\ \text{Var}(X) &= 11 - 3^2 = 2. \end{aligned}$$

4. Una variable aleatoria X toma solamente dos valores x_1 y x_2, tales que $x_2 > x_1$. La probabilidad de que X tome el valor x_1 es 0.6. Hallar la ley que sigue X, si la esperanza matemática y la varianza son conocidas: $\mathbf{E}(X) = 1.4$ y $\text{Var}(X) = 0.24$.

Solución: Escribamos la ley de X:

X	x_1	x_2
p	0.6	0.4

Expresemos la esperanza y la varianza en función de x_1 y x_2:

$$\mathbf{E}(X) = 0.6x_1 + 0.4x_2 = 1.4.$$

La ley de X^2 es

X^2	x_1^2	x_2^2
p	0.6	0.4

Entonces,

$$\mathbf{E}(X^2) = 0.6x_1^2 + 0.4x_2^2$$

y

$$\text{Var}(X) = \mathbf{E}(X^2) - [\mathbf{E}(X)]^2 = 0.6x_1^2 + 0.4x_2^2 - 1.4^2 = 0.24.$$

De aquí, se obtiene el sistema de ecuaciones

$$\begin{cases} 0.6x_1 + 0.4x_2 = 1.4 \\ 0.6x_1^2 + 0.4x_2^2 = 2.2 \end{cases}$$

Resolviendo el sistema se obtienen dos pares de soluciones:

$$x_1 = 1,\ x_2 = 2 \quad \text{y} \quad x_1 = 1.8,\ x_2 = 0.8.$$

Puesto que $x_2 > x_1$, hallamos la ley de distribución de X:

X	1	2
p	0.6	0.4

5. a) Hallar la esperanza y la varianza de una variable aleatoria X que tiene función de distribución:

$$F(x) = \begin{cases} 0, & \text{si } x \leq -1; \\ \dfrac{x+1}{4}, & \text{si } -1 < x \leq 3; \\ 1, & \text{si } x > 3. \end{cases}$$

b) Se define la variable aleatoria $Y = 5X + 2$. Hallar su esperanza y su varianza.

Solución:

a) Hallamos la función de densidad

$$f(x) = \begin{cases} \dfrac{1}{4}, & \text{si } -1 < x \leq 3; \\ 0, & \text{caso contrario.} \end{cases}$$

De manera que

$$\begin{aligned} \mathbf{E}(X) &= \int_{-1}^{3} xf(x)\,dx = \int_{-1}^{3} x\left(\frac{1}{4}\right)dx = \left.\frac{x^2}{8}\right|_{-1}^{3} = 1, \\ \mathbf{E}\left(X^2\right) &= \int_{-1}^{3} x^2 f(x)\,dx = \int_{-1}^{3} x^2\left(\frac{1}{4}\right)dx = \left.\frac{x^3}{12}\right|_{-1}^{3} = \frac{7}{3}. \end{aligned}$$

Por lo tanto,

$$\text{Var}(X) = \mathbf{E}\left(X^2\right) - (\mathbf{E}(X))^2 = \frac{7}{3} - 1^2 = \frac{4}{3}.$$

b) Tenemos la variable aleatoria $Y = 5X + 2$, cuya función de densidad no la conocemos, pero podemos emplear las propiedades de la esperanza y de la varianza:

$$\begin{aligned} \mathbf{E}(Y) &= \mathbf{E}(5X+2) = 5\mathbf{E}(X) + \mathbf{E}(2) = 5 \times 1 + 2 = 7, \\ \text{Var}(Y) &= \text{Var}(5X+2) = 25\,\text{Var}(X) = 25 \times \frac{4}{3} = \frac{100}{3}. \end{aligned}$$

6. Una variable aleatoria X está definida por su densidad $f(x) = x + \frac{1}{2}$ en el intervalo (0, 1), fuera de este intervalo $f(x) = 0$. Hallar la esperanza matemática de la variable aleatoria $Y = X^3$.

Solución:

$$\begin{aligned} \mathbf{E}(Y) &= \mathbf{E}(X^3) = \int_0^1 x^3 f(x)\,dx = \int_0^1 x^3\left(x + \frac{1}{2}\right)dx \\ &= \int_0^1 \left(x^4 + \frac{1}{2}x^3\right)dx = \left[\frac{x^5}{5} + \frac{1}{2}\cdot\frac{x^4}{4}\right]_0^1 \\ &= \frac{1}{5} + \frac{1}{2} \times \frac{1}{4} - 0 = \frac{13}{40}. \end{aligned}$$

7. Determinar la esperanza y la varianza de una variable aleatoria T cuya función de distribución es $F(t) = 1 - e^{-2t}$, $t > 0$.

Solución: La función de densidad es: $f(t) = F'(t) = 2e^{-2t}$, $t > 0$; y 0, caso contrario.

Calculemos la esperanza:

$$\mathbf{E}(X) = \int_{-\infty}^{\infty} xf(x)\,dx = \int_0^{\infty} 2xe^{-2x}dx.$$

Integrando por partes, poniendo $u = x$, $dv = e^{-2x}\,dx$; por lo tanto: $du = dx$, $v = -\frac{1}{2}e^{-2x}$,

$$\begin{aligned} \int_0^{\infty} xe^{-2x}dx &= \frac{-xe^{-2x}}{2}\bigg|_0^{\infty} + \frac{1}{2}\int_0^{\infty} e^{-2x}dx \\ &= -\frac{xe^{-2x}}{2}\bigg|_0^{\infty} - \frac{1}{4}e^{-2x}\bigg|_0^{\infty} \\ &= \frac{1}{2} - \frac{1}{4} = \frac{1}{4}. \end{aligned}$$

Entonces,

$$\mathbf{E}(X) = 2\int_0^{\infty} xe^{-2x}dx = 2\left(\frac{1}{4}\right) = \frac{1}{2}.$$

Necesitamos el cálculo de $\mathbf{E}(X^2)$:

$$\mathbf{E}(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx = 2\int_0^{\infty} x^2e^{-2x}dx.$$

Integrando por partes, dos veces, se llega a

$$2\int_0^{\infty} x^2e^{-2x}dx = \frac{2}{4} = \frac{1}{2}.$$

La varianza buscada es

$$\text{Var}(X) = \mathbf{E}\left(X^2\right) - (\mathbf{E}(X))^2 = \frac{1}{2} - \left(\frac{1}{2}\right)^2 = \frac{1}{4}.$$

8. En un supermercado se vende una cantidad variable de carne. Si X es la variable aleatoria que describe la cantidad de kilogramos que se vende al día, cuya función de densidad es:

$$f(x) = \begin{cases} \dfrac{x}{1250}, & \text{si } 0 \leq x \leq 50; \\ 0, & \text{caso contrario.} \end{cases}$$

a) ¿Cuál es la cantidad de carne que se espera vender diariamente en el supermercado? También, halle la desviación estándar.

b) Si la ganancia en el producto se expresa por la ecuación $G = \frac{2}{175}X^2 + 10$. Calcule la ganancia esperada.

Solución:

a) Calcularemos la esperanza de la variable aleatoria como el indicador de las ventas que efectúa el supermercado.

$$\begin{aligned} \mathbf{E}(X) &= \int_0^{50} x\left(\frac{x}{1250}\right) dx = \frac{1}{1250}\int_0^{50} x^2\, dx \\ &= \frac{1}{1250}\left[\frac{x^3}{3}\right]_0^{50} = 33.33. \end{aligned}$$

Así, el supermercado esperaría vender 33.33 kg diarios de carne.
Calculemos $\mathbf{E}\left(X^2\right)$:

$$\begin{aligned} \mathbf{E}\left(X^2\right) &= \int_0^{50} x^2\left(\frac{x}{1250}\right) dx = \frac{1}{1250}\int_0^{50} x^3\, dx \\ &= \frac{1}{1250}\left[\frac{x^4}{4}\right]_0^{50} = 1250. \end{aligned}$$

La varianza es

$$\begin{aligned} \mathrm{Var}(X) &= \mathbf{E}\left(X^2\right) - (\mathbf{E}(X))^2 = 1250 - (33.33)^2 \\ &= 138.89. \end{aligned}$$

La desviación estándar es $\sigma = \sqrt{\mathrm{Var}(X)} = \sqrt{138.89} = 11.79\,\mathrm{kg}$.

b) Entonces,

$$\begin{aligned} \mathbf{E}(G) &= \mathbf{E}\left(\frac{2}{175}X^2 + 10\right) = \frac{2}{175}\mathbf{E}\left(X^2\right) + 10 \\ &= \frac{2}{175}(1250) + 10 = 24.29. \end{aligned}$$

Por la venta de la carne, el supermercado espera ganar 24.29 dólares diarios.

A continuación, se da un ejemplo de variable aleatoria discreta que toma un número infinito de valores. El cálculo de la esperanza y de la varianza de este tipo de variables implica el uso de artificios que contemplan el empleo de elementos de análisis matemático, aquí se aplicará un método empleado por F. Samaniego[1] para las variables aleatorias geométricas, mediante propiedades básicas de las series geométricas:

[1] Samaniego, F. (1992), "Elementary Derivations of Geometric Moments," *The American Statistician*, 46, 108-109.

- Propiedad 1. $\forall a \in (0,1)$, $\sum\limits_{k=0}^{\infty} a^k = \dfrac{1}{1-a}$.
- Propiedad 2. $\forall a \in (0,1)$, $\sum\limits_{k=n}^{\infty} (1-a)a^k = a^n$.

9. Una variable aleatoria X toma valores 1, 2, 3, ..., con probabilidades

$$p_k = -\frac{1}{\ln p}\frac{(1-p)^k}{k}, \qquad k = 1, 2, \ldots; \quad p \in (0,1).$$

Determinar su esperanza y su varianza.

Solución:

$$\begin{aligned}\mathbf{E}(X) &= \sum_{k=1}^{\infty} p_k\, x_k = \sum_{k=1}^{\infty}\left(-\frac{1}{\ln p}\right)\frac{(1-p)^k}{k}\cdot k \\ &= \left(-\frac{1}{\ln p}\right)\sum_{k=1}^{\infty}(1-p)^k = -\left(\frac{1-p}{\ln p}\right)\sum_{k=0}^{\infty}(1-p)^k.\end{aligned}$$

Si $p \in (0,1)$, entonces $1-p \in (0,1)$ y aplicando la propiedad 1:

$$\mathbf{E}(X) = -\left(\frac{1-p}{\ln p}\right)\frac{1}{1-(1-p)} = -\left(\frac{1}{\ln p}\right)\left(\frac{1-p}{p}\right).$$

Para el cálculo de la varianza vamos a calcular $\mathbf{E}\left(X^2\right)$, usando la notación $q = 1-p$; por tanto,

$$\begin{aligned}\mathbf{E}\left(X^2\right) &= \sum_{k=1}^{\infty} p_k\, x_k^2 = \sum_{k=1}^{\infty}\left(-\frac{1}{\ln p}\right)\frac{q^k}{k}\cdot k^2 \\ &= \left(-\frac{1}{\ln p}\right)\sum_{k=1}^{\infty} kq^k = \left(-\frac{1}{\ln p}\right)\frac{1}{p}\sum_{k=1}^{\infty} kpq^k.\end{aligned}$$

Veamos a qué es igual la sumatoria:

$$\sum_{k=1}^{\infty} kpq^k = \sum_{k=1}^{\infty} pq^k + \sum_{k=2}^{\infty} pq^k + \sum_{k=3}^{\infty} pq^k + \cdots$$

Por la propiedad 2, se tiene que

$$\sum_{k=1}^{\infty} kpq^k = q + q^2 + q^3 + \cdots = \sum_{k=1}^{\infty} q^k = \frac{q}{p} = \frac{1-p}{p}.$$

Consecuentemente, $\mathbf{E}\left(X^2\right)$ queda como

$$\mathbf{E}\left(X^2\right) = -\left(\frac{1}{\ln p}\right)\left(\frac{1}{p}\right)\frac{1-p}{p} = -\frac{1-p}{p^2\ln p}.$$

Por lo tanto,

$$\begin{aligned}\mathrm{Var}(X) &= \mathbf{E}\left(X^2\right) - (\mathbf{E}(X))^2 = -\frac{1-p}{p^2\ln p} - \left(-\frac{1}{\ln p}\right)^2\left(\frac{1-p}{p}\right)^2 \\ &= (1-p)\frac{p-\ln p-1}{p^2\left(\ln^2 p\right)}.\end{aligned}$$

3.6. Función generadora de momentos

Los momentos de una variable aleatoria son números que representan algunas características de la distribución de probabilidad asociada. Bajo ciertas condiciones el conjunto de momentos determinan de manera única a la ley de probabilidad.

Definición (Momentos) Sea X una variable aleatoria y sea n un número natural. Cuando existe, el número $\mu_k = \mathrm{E}\left(X^k\right)$ es el $k-$ésimo momento de X.

Entonces, tenemos que la media μ es el primer momento de la variable aleatoria; es decir, $\mu = \mu_1$.

Asociada a cada variable aleatoria podemos encontrar una función que permite calcular sus momentos. Esta función tiene la propiedad de que, al igual que la función de distribución, caracteriza de manera única a la ley de probabilidad de la que proviene

Definición (Función generadora de momentos) La función generadora de momentos (f.g.m) de la variable aleatoria X es la función

$$M(t) = E\left(e^{tX}\right),$$

definida para valores reales de t tales que la esperanza existe.

La función generadora de momentos se utiliza tanto para variables aleatorias discretas como continuas.

Ejemplos

1. Encontrar la función generadora de momentos de la variable aleatoria cuya función probabilidad es

X	-4	6	10
p	0.2	0.3	0.5

Solución: Resulta que

$$\begin{aligned} M(t) &= E\left(e^{tX}\right) = \sum_k p_k e^{tk} \\ &= 0.2\,e^{-4t} + 0.3\,e^{6t} + 0.5\,e^{10t}. \end{aligned}$$

2. Hallar la función generadora de momentos de una variable aleatoria cuya función de densidad es

$$f(x) = \begin{cases} \frac{1}{4}, & \text{si } -1 < x \leq 3; \\ 0, & \text{caso contrario.} \end{cases}$$

Solución: Tenemos que

$$M(t) = E\left(e^{tX}\right) = \int_{-1}^{3} e^{tx} f(x)\,dx = \int_{-1}^{3} e^{tx}\frac{1}{4}dx = \frac{e^{3t} - e^{-t}}{4t}.$$

El siguiente resultado nos indica cómo se pueden obtener los momentos de cualquier orden con el empleo de la f.g.m.

Teorema. Sea X con función generadora de momentos $M(t)$, con derivadas continuas de cualquier orden; entonces,

$$\mu_k = E\left(X^k\right) = \frac{d}{dx^k}M(t)\Big|_{t=0}.$$

Es decir, el $k-$ésimo momento de una variable aleatoria se calcula como la derivada de orden k de la f.g.m., evaluada en cero.

Observación. Se tiene que $\mathbf{E}(X) = \mu_1$ y $\text{Var}(X) = \mu_2 - \mu_1^2$.

Ejemplos

1. Hallar la varianza de la variable aleatoria cuya función probabilidad es

X	-4	6	10
p	0.2	0.3	0.5

Solución: Antes calculamos que $M(t) = 0.2\,e^{-4t} + 0.3\,e^{6t} + 0.5\,e^{10t}$; por tanto,

$$\begin{aligned} \frac{d}{dt}M(t) &= -0.8e^{-4t} + 1.8e^{6t} + 5e^{10t} \\ \frac{d^2}{dt^2}M(t) &= 3.2e^{-4t} + 10.8e^{6t} + 50e^{10t}. \end{aligned}$$

Entonces,

$$\begin{aligned} \mu_1 &= \frac{d}{dt}M(0) = -0.8e^{-4(0)} + 1.8e^{6(0)} + 5e^{10(0)} = 6, \\ \mu_2 &= \frac{d^2}{dt^2}M(0) = 3.2e^{-4(0)} + 10.8e^{6(0)} + 50e^{10(0)} = 64. \end{aligned}$$

Por tanto, $\text{Var}(X) = \mu_2 - \mu_1^2 = 64 - (6)^2 = 28$.

2. Hallar la varianza de la variable aleatoria cuya función de densidad es

$$f(x) = \begin{cases} \frac{1}{4}, & \text{si } -1 < x \leq 3; \\ 0, & \text{caso contrario.} \end{cases}$$

Solución: Ya calculamos que

$$M(t) = \frac{e^{3t} - e^{-t}}{4t}.$$

Por tanto,

$$\begin{aligned} \frac{d}{dt}M(t) &= \frac{(3t-1)e^{3t} + (1+t)e^{-t}}{4t^2} \\ \frac{d^2}{dt^2}M(t) &= \frac{\left(9t^2 - 6t + 2\right)e^{3t} - \left(t^2 + 2t + 2\right)e^{-t}}{4t^3}. \end{aligned}$$

Puesto que $M(t)$ no está definida para $t = 0$, para hallar $\frac{d}{dt}M(0)$ y $\frac{d^2}{dt^2}M(0)$ aplicaremos la regla de L'Hospital; entonces,

$$\mu_1 = \frac{d}{dt}M(0) = \lim_{t\to 0} \frac{\frac{d^2}{dt^2}\left((3t-1)e^{3t} + (1+t)e^{-t}\right)}{\frac{d^2}{dt^2}(4t^2)} = 1,$$

$$\mu_2 = \frac{d^2}{dt^2}M(0) = \lim_{t\to 0} \frac{\frac{d^3}{dt^3}\left(\left(9t^2 - 6t + 2\right)e^{3t} - \left(t^2 + 2t + 2\right)e^{t}\right)}{\frac{d^3}{dt^3}(4t^3)} = \frac{56}{24}.$$

Por lo tanto,

$$\mathrm{Var}(X) = \mu_2 - \mu_1^2 = \frac{56}{24} - 1^2 = \frac{4}{3}.$$

3.7. Ejercicios

1. Halle la esperanza y la varianza de las variables aleatorias discretas definidas por

X	-0.71	0.24	0.61
p	0.2	0.5	0.3

Y	2	4	5	6
p	0.3	0.1	0.2	0.4

2. Se escoge aleatoriamente un número de conjunto $S = \{-1; 0; 1\}$. Sea X el número escogido. Encuentre su valor esperado, la varianza y la desviación estándar de X.

3. ¿Existe una variable aleatoria X que cumple que $\mathbf{E}(X-2) = 8$ y que $\mathbf{E}\left((X+1)^2\right) = 120$?

4. Si X es una variable aleatoria con $\mathbf{E}(X) = 50$ y $\mathrm{Var}(X) = 15$. Encuentre:

   a) $\mathbf{E}\left(X^2\right)$;
   b) $\mathbf{E}(-X)$;
   c) $\mathrm{Var}(-X)$;
   d) $\sigma(-2X)$;
   e) $\mathbf{E}(3X+10)$;
   f) $\mathrm{Var}(3X+10)$.

5. Sean X, Y y Z tres variables aleatorias independientes, cada una con media μ y varianza σ^2.

   a) Encuentre la esperanza y la varianza de $S = X + Y + Z$;
   b) Encuentre la esperanza y la varianza de $T = 3X$;
   c) Calcule la esperanza y la varianza de $A = \dfrac{X+Y+Z}{3}$;
   d) Encuentre la esperanza de S^2 y de A^2.

6. Sea X una variable aleatoria definida en el intervalo $[-1; 1]$ y sea $f_X(x)$ su función de densidad. Encuentre $\mathbf{E}(X)$ y $\mathrm{Var}(X)$ si:

   a) $f_X(x) = \frac{1}{2}$;
   b) $f_X(x) = |x|$.
   c) $f_X(x) = 1 - |x|$.
   d) $f_X(x) = \frac{3}{2}x^2$.

7. Encuentre la esperanza y desviación estándar de las variables aleatorias definidas mediante las leyes:

   a) $$F(x) = \begin{cases} 0, & \text{si } x \le 1; \\ \dfrac{x-1}{3}, & \text{si } 1 < x \le 4; \\ 1, & \text{si } x > 4. \end{cases}$$

b) $F(x) = \begin{cases} 0, & \text{si } x \leq 1; \\ \dfrac{x^2 - x}{2}, & \text{si } 1 < x \leq 2; \\ 1, & \text{si } x > 2. \end{cases}$

c) $f(x) = \begin{cases} 2x, & \text{si } x \in (0,\, 1); \\ 0, & \text{si } x \notin (0,\, 1). \end{cases}$

d) $f(x) = \begin{cases} -\dfrac{5}{22}(x^2 + 8x + 12), & \text{si } x \in (-5, -3); \\ 0, & \text{caso contrario.} \end{cases}$

e) $f(x) = \begin{cases} \dfrac{c}{x}, & \text{si } x \in [1,\, 2]; \\ 0, & \text{si } x \notin [1,\, 2]. \end{cases}$

(Determine primero el valor de c.)

8. Una variable aleatoria X toma los valores 4, 6 y α con probabilidades $\Pr(X = 4) = 0.5$, $\Pr(X = 6) = 0.3$ y $\Pr(X = \alpha) = p$. Si se sabe que la esperanza de X es igual a 6, halle los valores de p y α.

9. Halle la varianza de una variable aleatoria Z que solo puede tomar dos valores, el uno el doble del otro, con la misma probabilidad, si se sabe que $\mathbf{E}(Z) = 0.9$.

10. La variable aleatoria discreta X tiene solamente dos posibles valores: x_1 y x_2, además $x_1 < x_2$. La probabilidad de que X tome el valor x_1 es igual a 0.2. Halle la ley de distribución de X, conociendo la esperanza $\mathbf{E}(X) = 2.6$ y la desviación estándar $\sigma = 0.8$.

11. Una variable aleatoria X puede tomar tres valores: $x_1 = -1$, $x_2 = 0$ y $x_3 = 1$. Si se conocen las esperanzas matemáticas $\mathbf{E}(X) = 0.1$ y $\mathbf{E}(X^2) = 0.8$, encuentre las probabilidades p_1, p_2 y p_3, de los valores x_1, x_2 y x_3, respectivamente.

12. La variable aleatoria X tiene únicamente tres posibles valores $x_1 = 1$, x_2 y x_3 $(x_1 < x_2 < x_3)$. Las probabilidades de que X tome los valores x_1 y x_2 son respectivamente iguales a 0.3 y 0.2. Determine la ley de distribución de X, conociendo la esperanza $\mathbf{E}(X) = 2.2$ y la varianza $\mathrm{Var}(X) = 0.76$.

13. La variable aleatoria X tiene función de distribución

$$F(x) = \begin{cases} 0, & \text{para } x \leq 2; \\ ax + b, & \text{para } 2 < x \leq 4; \\ 1, & \text{para } x > 4. \end{cases}$$

a) Halle el valor de las constantes a y b;

b) Determine su función de densidad;

c) Halle las probabilidades: $\Pr(1 < X < 3)$ y $\Pr(X > 2.5)$;

d) Encuentre su esperanza y su varianza.

14. Suponga que se escoge un número real X en el intervalo $[2; 10]$ con una función de densidad de la forma $f(x) = Cx$, donde C es una constante.

a) Halle el valor de C;

b) Calcule $\Pr(D)$, donde $D = [3; 7]$;

c) Encuentre $\Pr(X > 5)$, $\Pr(X < 7)$ y $\Pr(X^2 - 12X + 35 > 0)$;

d) Encuentre la esperanza y la varianza de X.

15. Un estudiante rinde una prueba consistente en 2 problemas de elección múltiple. La primera tiene 3 posibles respuestas y la segunda 5. El estudiante escoge las 2 respuestas al azar. Encuentre:

   a) La ley que describe el número de respuestas correctas X del estudiante;
   b) El número esperado, $\mathbf{E}(X)$, de respuestas correctas;
   c) La varianza, $\text{Var}(X)$.

16. Una organización benéfica realiza una rifa para conseguir fondos, en la que se vendieron 10 000 boletos, a 4 dólares cada uno. El premio es un automóvil de 12 000 dólares. Si un ciudadano compra 2 boletos, ¿cuál es la ganancia esperada del comprador de los boletos?

17. Una persona participa en un concurso de la televisión. Le hacen una pregunta con 5 respuestas (solo una es verdadera) si acierta, gana 10 000. Si falla le vuelven hacer otra pregunta con tres posibles respuestas de las cuales solo una es verdadera. Si acierta, gana 1000 y si falla se le vuelve hacer otra pregunta con solo dos respuestas si acierta, entonces no gana nada y si falla pierde 500. El juego termina cuando la persona acierta o después de fallar la tercera pregunta.

   a) Indique cuál es el espacio muestral;
   b) Calcule la probabilidad de que dé una respuesta correcta;
   c) Halle la ganancia esperada;
   d) Halle la esperanza y la varianza de la variable aleatoria que describe el número de preguntas realizadas al concursante.

18. Se asegura un vehículo de 50 000 dólares por su valor total, pagando una prima de C. Si la probabilidad de robo en un año es de 0.02, ¿cuál es el valor de la prima que debe cobrar la compañía de seguros, si espera ganar 200 dólares?

19. Si Roberto termina sus estudios en Junio, podrá disfrutar de una beca para poder realizar un curso de especialización con todos los gastos pagados. Si aprueba en Septiembre, la beca sólo le cubrirá el 40 % de los gastos. Si no consigue aprobar, también realizará el curso pero abonando 50 000 dólares, que es lo que cuesta. Roberto sabe que la probabilidad de aprobar en Junio es sólo de un 10 %, mientras que la de aprobar en Septiembre es de un 40 %.

   a) Halle la distribución de probabilidad del costo del curso de especialización;
   b) Determine el valor esperado de dicha variable.

20. Una agencia que renta autos tiene disponibles 4 carros todo terreno, para alquilarlos. El precio de alquiler de cada carro es 60 dólares diarios. En un estudio de mercado el propietario ha determinado el siguiente modelo probabilístico sobre la demanda de estos autos:

Demanda	**Probabilidad**
0	0.05
1	0.10
2	0.20
3	0.25
4	0.20
5	0.15
6	0.05

Además, en el mismo estudio ha encontrado que sus gastos diarios son: 20 dólares por alquiler del local y 15 por pago a un empleado.

a) Calcule el número esperado de carros todo terreno que la agencia alquilará un día cualquiera;

b) Calcule la ganancia diaria esperada;

c) Calcule la desviación estándar de la ganancia.

21. Un portafolio de inversión sigue el siguiente esquema probabilístico:

Compañía	Ganancia	Probabilidad	Compañía	Ganancia	Probabilidad
A	−200	0.0102	F	250	0.2123
B	−150	0.0346	G	350	0.1542
C	−50	0.0860	H	450	0.0860
D	50	0.1542	I	550	0.0346
E	150	0.2123	J	650	0.0156

Calcule la esperanza y la desviación estándar de la ganancia de una inversión en este portafolio.

22. Un círculo de radio 1 es zonificado en 10 círculos concéntricos de radios $\frac{1}{10}, \frac{2}{10}, \ldots, \frac{10}{10} = 1$. Se lanza un dardo sobre el círculo, si éste cae en la zona comprendida entre los círculos de radios $\frac{i}{10}$ e $\frac{i+1}{10}$ el lanzador gana $10 - i$ dólares, $i = 0, 1, \ldots, 9$. Sea X la cantidad de dinero ganado,

a) Halle la ley de la variable aleatoria X;

b) Calcule su esperanza y su varianza.

23. El espesor del recubrimiento de unos cables tiene función de densidad $\frac{600}{x^2}$, con $100\,\mu\text{m} < x < 200\,\mu\text{m}$.

a) Determine la media y la varianza del espesor del recubrimiento;

b) Si el costo del recubrimiento es de 0.5 dólares por micrómetro de espesor, ¿cuál es el costo medio por recubrir los cables?

24. Un supermercado tiene una demanda diaria variable X de la cantidad de carne que vende, de tal manera que X (medida en cientos de kilogramos) tiene una función de densidad

$$f(x) = \begin{cases} \frac{3}{64}x^2, & \text{si } 0 \leq x < 4; \\ 0, & \text{caso contrario.} \end{cases}$$

a) Calcule el valor esperado de la cantidad de carne demandada y su desviación estándar;

b) Si la ganancia está dada por $Y = 2X - 0.5$. Calcule la ganancia esperada y su varianza.

25. La longitud de ciertos gusanos se distribuye según la función de densidad

$$f(x) = \begin{cases} \frac{3}{4}(x-1)(3-x), & \text{si } 1 \leq x < 3; \\ 0, & \text{caso contrario.} \end{cases}$$

a) Calcule la esperanza y la varianza de la longitud de los gusanos;

b) Si para un estudio se considerarán aquellos ejemplares que tengan una longitud entre 1.7 cm y 2.4 cm, calcule el porcentaje de gusanos que tienen esta característica.

26. El tiempo de uso diario de la red Internet en una oficina tiene por función de densidad (medida en horas) a

$$f(x) = \begin{cases} \frac{3x^2(8-x)}{1024}, & \text{si } 0 \leq x \leq 8; \\ 0, & \text{caso contrario.} \end{cases}$$

a) Calcule el valor esperado y la varianza del tiempo diario de uso de la red Internet.

b) El tiempo de uso de Internet cuesta 2 dólares por hora. Calcule el valor esperado y la desviación estándar del costo semanal (en 5 días laborables) por el citado uso.

27. La ley de probabilidad que describe la distancia (en metros) a la que un atleta lanza la jabalina es

$$f(x) = \begin{cases} \dfrac{1}{20\pi}\,\text{sen}\left(\dfrac{x}{10\pi}\right), & \text{si } x \in [0, 10\pi^2]; \\ 0, & \text{caso contrario.} \end{cases}$$

a) Halle la probabilidad de que una jabalina lanzada llegue a una distancia mayor que 60 m;

b) Determine el valor esperado de la distancia a la que llega la jabalina;

c) Halle la varianza y la desviación estándar de la distancia cubierta por la jabalina.

28. Demuestre que la esperanza y la varianza de la variable aleatoria discreta definida por

$$\Pr(X = k) = pq^k, \quad k = 0, 1, 2, \ldots$$

con $q = 1 - p$, vienen dadas por $\mathbf{E}(X) = \dfrac{q}{p}$ y $\text{Var}(X) = \dfrac{q}{p^2}$.

29. Sea X una variable aleatoria discreta que puede tomar valores enteros, definida por

$$\Pr(X = x) = \begin{cases} \alpha, & \text{para } x = -1; \\ (1-\alpha)^2\alpha^x, & \text{para } x = 0, 1, 2, \ldots \end{cases}$$

para algún α en $(0, 1)$.

a) Pruebe que la función de probabilidad está bien definida.

b) Demuestre que $\mathbf{E}(X) = 0$.

30. Sean X y Y dos variables aleatorias independientes. Si X toma los valores 3, 1 y 2 con probabilidades que son los 3 términos de una progresión aritmética cuya diferencia es $\dfrac{1}{4}$; mientras tanto, Y toma los valores 2, 1 y 0 con probabilidades que forman una progresión geométrica de razón $\dfrac{1}{2}$. Calcule la esperanza de $XY + 2Y - X$.

31. Dadas las variables aleatorias independientes X y Y, cuyas funciones de densidad son:

$$f_X(x) = \begin{cases} 2x, & \text{si } x \in [0,\ 1]; \\ 0, & \text{caso contrario.} \end{cases} \qquad f_Y(y) = \begin{cases} \dfrac{y}{2}, & \text{si } x \in [0,\ 2]; \\ 0, & \text{caso contrario.} \end{cases}$$

Calcule:

a) $\mathbf{E}(X + Y)$; b) $\mathbf{E}(2X - 3Y + 5)$; c) $\text{Var}(4X - 2Y - 3)$.

32. Las variables aleatorias $X_1, X_2, \ldots, X_n$ son independientes e igualmente distribuidas, tales que

$$\Pr(X_i = 1) = \Pr(X_i = -1) = \frac{1}{4},$$

$$\Pr(X_i = 0) = \frac{1}{2}, \qquad i = 1, 2, \ldots, n.$$

Halle la esperanza matemática y la varianza de la variable aleatoria $S_n = X_1 + X_2 + \cdots + X_n$.

33. Las variables aleatorias $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_n$ son independientes. Pongamos

$$S_n = X_1 + X_2 + \cdots + X_n,$$

$$T_n = X_1Y_1 + X_2Y_2 + \cdots + X_nY_n.$$

Halle $\mathbf{E}(S_n)$, $\mathbf{E}(T_n)$, $\mathrm{Var}(S_n)$ y $\mathrm{Var}(T_n)$ si

$$\mathbf{E}(X_k) = a, \quad \mathrm{Var}(X_k) = \sigma^2,$$

$$\Pr(Y_k = 1) = p, \quad \Pr(Y_k = 0) = q = 1 - p, \quad k = 1, 2, \ldots, n.$$

34. Para las variables aleatorias que tienen las siguientes leyes, encuentre la función generadora de momentos y, mediante esta función, halle la esperanza y la varianza:

a)

Y	1	2	3	4	5
p	1/5	1/5	1/5	1/5	1/5

b)

Y	2	4	5	6
p	0.3	0.1	0.2	0.4

c)

X	0	2	−2
p	0.2	0.5	0.3

d) $f(x) = \begin{cases} \dfrac{1}{6}, & \text{si } x \in (-2, 4); \\ 0, & \text{si } x \notin (-2, 4). \end{cases}$

e) $f(x) = \begin{cases} e^{-3x}, & \text{si } x \geq 0; \\ 0, & \text{si } x < 0. \end{cases}$

f) $f(x) = \begin{cases} 3 \operatorname{sen} 3x, & \text{si } \dfrac{\pi}{6} < x \leq \dfrac{\pi}{3}; \\ 0, & \text{caso contrario.} \end{cases}$

35. Sean X y Y dos variables aleatorias independientes e idénticamente distribuidas con f.g.m. $M(t)$. Demuestre que $M_{X+Y}(t) = M_X(t)M_Y(t)$ y que $M_{X-Y}(t) = M_X(t)M_Y(-t)$.

36. Sea X con f.g.m. $M_X(t)$, y sean a y b dos constantes. Demuestre que $M_{aX+b}(t) = e^{tb}M_X(at)$.

Capítulo 4

Principales Distribuciones de Probabilidad

Hay que explorar sistemáticamente el azar.
Facultad de Letras, París.

En este capítulo se presentan, en detalle, algunos tipos de leyes que siguen las variables aleatorias, que aparecen frecuentemente en problemas prácticos y cuyas propiedades deben ser conocidas.

A una variable aleatoria X que sigue una ley $\mathcal{L}$ de parámetros (p_1, p_2) la notaremos como $X \sim \mathcal{L}(p_1, p_2)$.

4.1. Distribución uniforme discreta

Una variable aleatoria X, que puede tomar un número finito de valores, 1, 2, ..., n, cada uno de los cuales tiene la misma probabilidad de ocurrir, se dice que sigue una ley de distribución *uniforme discreta*. Es decir,

$$\Pr(X = k) = \frac{1}{n}, \quad k = 1, 2, \ldots, n.$$

Su esperanza es igual a

$$\mathbf{E}(X) = \frac{n+1}{2}$$

y su varianza

$$\mathrm{Var}(X) = \frac{n^2 - 1}{12}.$$

La equiprobabilidad es la forma más obvia de asignar probabilidades dentro de un fenómeno aleatorio cuyo comportamiento es desconocido. Esta ley aparece en los juegos de azar en los que todos los jugadores tienen iguales posibilidades; además, esta ley es la básica en la simulación de eventos aleatorios mediante computadora.

Ejemplos

1. Se lanza un dado en forma de octaedro, considerando la variable aleatoria que describe el número de puntos que aparece. Determinar su esperanza y varianza.

Solución: Se tiene que $n = 8$ y se asigna la probabilidad $\Pr(X = i) = \frac{1}{8}$; entonces,

$$\begin{aligned} \mathbf{E}(X) &= \frac{n+1}{2} = \frac{8+1}{2} = \frac{9}{2}. \\ \mathrm{Var}(X) &= \frac{n^2-1}{12} = \frac{64-1}{12} = \frac{21}{4}. \end{aligned}$$

2. Una máquina registra, en minutos completos, la diferencia de tiempo en el paso de camiones por cierto lugar de la carretera. Se sabe que la diferencia máxima puede ser 9 minutos. Si se asume que los arribos son aleatorios, calcular el tiempo que se esperaría exista entre dos arribos consecutivos, su varianza y desviación estándar.

   Solución: La variable aleatoria puede tomar los valores 1, 2, ..., 9, que suponemos tiene distribución uniforme, por lo tanto,

$$\begin{aligned} \mathbf{E}(X) &= \frac{n+1}{2} = \frac{9+1}{2} = 5\,\mathrm{min}, \\ \mathrm{Var}(X) &= \frac{n^2-1}{12} = \frac{81-1}{12} = 6.67\,\mathrm{min}^2, \\ \sigma &= \sqrt{6.67} = 2.58\,\mathrm{min}. \end{aligned}$$

3. Un reloj está descompuesto y suena, aleatoriamente, a la hora en punto; es decir, puede sonar a la una, a las dos, ..., a las doce; dando ese mismo número de campanadas. Determinar la esperanza y varianza de la variable aleatoria que describe el número de campanadas que se habrá de esperar que dé el reloj.

   Solución: Encontremos la esperanza y la varianza, considerando $n = 12$:

$$\begin{aligned} \mathbf{E}(X) &= \frac{n+1}{2} = \frac{12+1}{2} = 6.5\,\mathrm{h}, \\ \mathrm{Var}(X) &= \frac{n^2-1}{12} = \frac{144-1}{12} = 11.92\,\mathrm{h}^2. \end{aligned}$$

4.2. Distribución hipergeométrica

Nos planteamos el siguiente problema: en una urna se tienen N bolas, n de las cuales son rojas y las $N - n$ restantes negras, de las cuales se extraen al azar r bolas; investigaremos la probabilidad de que el grupo elegido contenga exactamente k bolas rojas. Aquí, k puede ser cualquier entero entre cero y n o r.

La probabilidad es

$$\Pr(X = k) = \frac{\mathbf{C}_n^k \mathbf{C}_{N-n}^{r-k}}{\mathbf{C}_N^r}, \quad k = 0, 1, \ldots, \min\{n, r\}.$$

Si consideramos la proporción de bolas rojas en la composición inicial de bolas contenidas en la urna, $p = \frac{n}{N}$ y $q = 1 - p$, la fórmula de la probabilidad puede expresarse como

$$\Pr(X = k) = \frac{\mathbf{C}_{Np}^k \mathbf{C}_{Nq}^{r-k}}{\mathbf{C}_N^r}, \quad k = 0, 1, \ldots, \min\{Np, r\},$$

por lo que la probabilidad p de obtener una bola roja se puede introducir como un parámetro que define la ley.

A una variable aleatoria X que sigue una ley hipergeométrica de parámetros (N, n, r) se la nota mediante $\mathcal{H}(N, n, r)$.

La esperanza es

$$\mathbf{E}(X) = \frac{rn}{N} = rp.$$

y la varianza

$$\text{Var}(X) = r\frac{n}{N}\left(1 - \frac{n}{N}\right)\left(\frac{N-r}{N-1}\right) = rp(1-p)\left(\frac{N-r}{N-1}\right).$$

Esta distribución de probabilidad surge en el análisis de muestras en control de la calidad de lotes de productos (en los cuales hay artículos útiles y defectuosos), en estudios censales de poblaciones animales y al realizar muestreo sin reposición.

La siguiente fórmula de recurrencia facilita el cálculo iterativo de las probabilidades de la ley[1]:

$$\begin{aligned} p_0 &= \frac{(N-n)!\,(N-r)!}{N!\,(N-n-r)!} \\ p_k &= p_{k-1}\frac{n+1-k}{k}\,\frac{r+1-k}{N-n-r+k}, \quad k = 1, 2, \ldots, n. \end{aligned}$$

Ejemplos

1. En un grupo de 12 estudiantes 8 son sobresalientes. Por lista se escogieron 9 al azar. a) ¿Cuál es la probabilidad de que entre los estudiantes seleccionados hayan 5 sobresalientes?; b) ¿Cuántos estudiantes sobresalientes se espera encontrar entre los seleccionados?

   Solución:

   a) Si consideramos la variable aleatoria Z: «Número de estudiantes sobresalientes entre los 9 escogidos», se tiene:

   Total de estudiantes, $N = 12$;

   Total de estudiantes sobresalientes, $n = 8$;

   Total de estudiantes escogidos, $r = 9$;

   Número de estudiantes sobresalientes escogidos, $k = 5$.

   Por lo tanto, $Z \sim \mathcal{H}(12, 8, 9)$ y

$$\Pr(Z = 5) = \frac{\mathbf{C}_8^5\mathbf{C}_4^4}{\mathbf{C}_{12}^9} = \frac{\dfrac{8!}{5!3!}\dfrac{4!}{4!0!}}{\dfrac{12!}{9!3!}} = \frac{14}{55}.$$

   b) Calculemos la esperanza de Z:

$$\mathbf{E}(Z) = \frac{rn}{N} = \frac{9 \times 8}{12} = 6$$

   Se esperaría encontrar 6 sobresalientes.

[1] Drane, J. W., Cao, S., Wang, L. y Postelnicu, T. (1993), "Limiting Forms of Probability Mass Functions via Recurrence Formulas," *The American Statistician*, 47, 269-274.

2. En un control de calidad industrial se toma un lote de 10 piezas para una inspección. En el lote hay 8 piezas correctas. Se toman al azar 2 piezas. Formar la ley de distribución del número de piezas correctas entre las escogidas y calcular su esperanza.

   Solución: La variable aleatoria Y (número de piezas correctas entre las escogidas) tiene los siguientes valores: $x_1 = 0$, $x_2 = 1$, $x_3 = 2$.

   Emplearemos la ley hipergeométrica con $N = 10$ (número total de piezas), $n = 8$ (número total de piezas correctas) y $r = 2$ (tamaño de la muestra); es decir, $Y \sim \mathcal{H}(10, 8, 2)$, obteniendo:

$$\Pr(Y = 0) = \frac{\mathbf{C}_8^0 \mathbf{C}_2^2}{\mathbf{C}_{10}^2} = \frac{1}{\dfrac{10 \times 9}{1 \times 2}} = \frac{1}{45},$$

$$\Pr(Y = 1) = \frac{\mathbf{C}_8^1 \mathbf{C}_2^1}{\mathbf{C}_{10}^2} = \frac{8 \times 2}{45} = \frac{16}{45},$$

$$\Pr(Y = 2) = \frac{\mathbf{C}_8^2 \mathbf{C}_2^0}{\mathbf{C}_{10}^2} = \frac{\dfrac{8 \times 7}{1 \times 2}}{45} = \frac{28}{45}.$$

   Formemos la ley de distribución buscada:

Y	0	1	2
p	1/45	16/45	28/45

   La esperanza es

$$\mathbf{E}(Y) = 0 \times \frac{1}{45} + 1 \times \frac{16}{45} + 2 \times \frac{28}{45} = 1.6.$$

4.3. Distribuciones de Bernoulli y binomial

Si al realizar un experimento una vez, solo hay dos resultados posibles, se tiene una prueba de Bernoulli. Se acostumbra referirse a uno de los resultados como «*éxito*», que aparece con probabilidad p, y al otro resultado como «*fracaso*», que sucede con probabilidad q. Es evidente que p y q son no negativos y que $p + q = 1$.

Generalmente, se define la variable aleatoria que sigue una ley de Bernoulli así:

$$X = \begin{cases} 1, & \text{si es éxito;} \\ 0, & \text{si es fracaso.} \end{cases}$$

La ley de probabilidad es

$$\Pr(X = 1) = p, \qquad \Pr(X = 0) = 1 - p = q.$$

La esperanza y la varianza son iguales a:

$$\mathbf{E}(X) = p, \qquad \mathrm{Var}(X) = pq.$$

La ley de Bernoulli desempeña un papel fundamental en el análisis de fenómenos en los cuales solo se tienen dos resultados mutuamente excluyentes, como es el caso de muchas preguntas en todo tipo de encuestas o la determinación del sexo de los recién nacidos.

Ejemplos

1. El más conocido es el del lanzamiento de una moneda homogénea, aquí $p = q = \frac{1}{2}$. Si hay desequilibrio en la moneda p y q son distintos de $\frac{1}{2}$.

2. Considérese el experimento consistente en lanzar un dado y la variable aleatoria X: «el número de puntos es mayor que 4». Entonces,

$$X = \begin{cases} 1, & \text{si } \omega \in \{5,6\}; \\ 0, & \text{si } \omega \in \{1,2,3,4\}. \end{cases}$$

y las probabilidades son

$$\Pr(X=0) = \frac{2}{3}, \qquad \Pr(X=1) = \frac{1}{3}.$$

3. Según datos censales se ha establecido que en la población ecuatoriana el 52 % lo constituyen las mujeres y el 48 % restante los hombres. Si se toma una persona al azar y se quiere conocer su sexo, se tiene

$$X = \begin{cases} 1, & \text{si es mujer;} \\ 0, & \text{si es hombre.} \end{cases}$$

y

$$\Pr(X=0) = 0.48, \qquad \Pr(X=1) = 0.52. \quad ◀$$

Supongamos que se realiza una sucesión de n pruebas de Bernoulli e interesa conocer el número de «éxitos» obtenidos, al margen del orden en que ellos se presenten. El número de éxitos puede ser 0, 1, 2, ..., n.

Se llama *binomial* a la ley de distribución de una variable aleatoria discreta X que describe el número k de éxitos en una sucesión de n pruebas de Bernoulli independientes, en cada una de las cuales la probabilidad de éxito es igual a p.

La ley de distribución binomial fue descubierta por James Bernoulli, quien la dejó escrita en su obra *Ars Conjectandi*, publicada en 1713, después de su muerte ocurrida en 1705.

La probabilidad se calcula mediante

$$\Pr(X=k) = \mathbf{C}_n^k p^k q^{n-k}, \quad k = 0, 1, \ldots, n.$$

A la variable aleatoria X que sigue una ley binomial de parámetros n y p se la notará como $X \sim \mathrm{Bin}(n,p)$.

Si $X \sim \mathbf{Bin}(n,p)$, se tiene que

$$\mathbf{E}(X) = np, \qquad \mathrm{Var}(X) = npq.$$

La distribución binomial tiene amplia aplicación en la teoría de muestreo cuando se puede contestar a una pregunta únicamente con dos opciones (por ejemplo SI-NO).

El cálculo de las probabilidades puede ser un proceso difícil porque los factoriales en los coeficientes binomiales crecen muy rápido, mientras que las potencias de p y q decrecen rápidamente. Por estas razones se utiliza la siguiente fórmula recursiva para su cálculo[2]:

$$\begin{aligned} p_0 &= (1-p)^n \\ p_k &= p_{k-1}\frac{n+1-k}{k}\frac{p}{1-p}, \quad k = 1, 2, \ldots, n. \end{aligned}$$

[2]Drane, J. W. y otros (1994), op. cit.

Ejemplos

1. Supongamos que en una población existen igual número de hombres y de mujeres y consideremos aquellas familias que tienen 4 hijos.

   a) Formar la ley de la variable aleatoria que describe el número de hijos varones en dichas familias.

   b) Calcular la probabilidad de que en una de estas familias haya más de un hijo varón.

   c) ¿Cuántos hijos varones se espera que haya en una familia que tiene 4 hijos?

Solución: Sabemos que $p = \dfrac{1}{2}$ y el número total de hijos es $n = 4$. Entonces, la variable aleatoria X: «Número de hijos varones», sigue una ley binomial de parámetros $(4, 1/2)$; o sea, $X \sim \mathbf{Bin}(4, 1/2)$.

a) Determinemos la probabilidades correspondientes.

- Si en la familia no hay ningún varón, $k = 0$:

$$\Pr(X = 0) = \mathbf{C}_4^0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^{4-0} = \frac{1}{16}.$$

- Si hay un varón, $k = 1$:

$$\Pr(X = 1) = \mathbf{C}_4^1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^{4-1} = \frac{1}{4}.$$

- Si hay dos varones, $k = 2$:

$$\Pr(X = 2) = \mathbf{C}_4^2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^{4-2} = \frac{3}{8}.$$

- Si hay tres varones, $k = 3$:

$$\Pr(X = 3) = \mathbf{C}_4^3 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^{4-3} = \frac{1}{4}.$$

- Si hay cuatro varones, $k = 4$:

$$\Pr(X = 4) = \mathbf{C}_4^4 \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^{4-4} = \frac{1}{16}.$$

De manera que

X	0	1	2	3	4
p	1/16	1/4	3/8	1/4	1/16

b) Debemos calcular $\Pr(X > 1) = \Pr(X = 2) + \Pr(X = 3) + \Pr(X = 4)$. Esta probabilidad, también se puede calcular mediante el evento complementario:

$$\begin{aligned}\Pr(X > 1) &= 1 - \Pr(X \leq 1) = 1 - [\Pr(X = 0) + \Pr(X = 1)] \\ &= 1 - \left(\frac{1}{16} + \frac{1}{4}\right) = \frac{11}{16}.\end{aligned}$$

c) El número esperado de hijos varones se calcula mediante la esperanza de la variable aleatoria:

$$\mathbf{E}(X) = np = 4 \times \frac{1}{2} = 2.$$

De manera que se esperaría que en una familia que tiene cuatro hijos, dos sean varones.

2. Un dispositivo está compuesto por tres elementos que trabajan independientemente. La probabilidad de falla de cada elemento en un día es igual a 0.1. Formar la ley de distribución del número de elementos que fallan en un día.

Solución: La variable aleatoria X («número de elementos que fallan») puede tomar los siguientes valores:

$x_1 = 0$ (ningún elemento falló), $x_2 = 1$ (falló un solo elemento), $x_3 = 2$ (fallaron dos elementos), $x_4 = 3$ (fallaron tres elementos).

Las probabilidades de fallo de cada uno de los elementos son iguales entre si, entonces es aplicable la ley binomial; por lo tanto, $X \sim \mathbf{Bin}(3, 0.1)$:

$$\begin{aligned} p_1 &= \Pr(X = 0) = \mathbf{C}_3^0(0.1)^0(0.9)^3 = 1 \cdot (0.1)^0 \cdot (0.9)^3 = 0.729, \\ p_2 &= \Pr(X = 1) = \mathbf{C}_3^1(0.1)^1(0.9)^2 = 3 \cdot 0.1 \cdot (0.9)^2 = 0.243, \\ p_3 &= \Pr(X = 2) = \mathbf{C}_3^2(0.1)^2(0.9)^1 = 3 \cdot (0.1)^2 \cdot 0.9 = 0.027, \\ p_4 &= \Pr(X = 3) = \mathbf{C}_3^3(0.1)^3(0.9)^0 = 1 \cdot (0.1)^3 \cdot (0.9)^0 = 0.001. \end{aligned}$$

En resumen,

X	0	1	2	3
p	0.729	0.243	0.027	0.001

3. Un examen consta de ocho preguntas de elección múltiple, cada una de ellas ofrece cinco alternativas, de las cuales solo una es correcta. Para aprobar el examen es necesario contestar correctamente al menos tres preguntas. Si un estudiante se propone responder a las preguntas al azar.

a) ¿Cuál es la probabilidad de que conteste correctamente todas las preguntas?;

b) ¿Cuál es la probabilidad de que el estudiante apruebe el examen?

Solución: La probabilidad de responder correctamente a cualquier pregunta es $p = \dfrac{1}{5} = 0.2$.

Sea Z: «número de preguntas correctamente contestadas», con $Z \sim \mathbf{Bin}(8, 0.2)$.

a) Debemos determinar la probabilidad del evento $Z = 8$:

$$\Pr(Z = 8) = \mathbf{C}_8^8(0.2)^8(0.8)^{8-8} = 1 \cdot (0.2)^8 \cdot 1 = 0.00000256.$$

Lo que nos indica que es muy difícil que adivine todas las respuestas.

b) Para aprobar se debe contestar correctamente al menos tres preguntas, por lo tanto $Z \geq 3$.

$$\begin{aligned} \Pr(Z \geq 3) &= 1 - \Pr(Z < 3) \\ &= 1 - [\Pr(Z = 0) + \Pr(Z = 1) + \Pr(Z = 2)] \\ &= 1 - \mathbf{C}_8^0(0.2)^0(0.8)^8 - \mathbf{C}_8^1(0.2)^1(0.8)^7 - \mathbf{C}_8^2(0.2)^2(0.8)^6 \\ &= 0.20308. \end{aligned}$$

4. Una agencia de turismo ofrece viajes a la amazonía. La utilidad mínima que le reporta uno de estos viajes es 6 dólares por cliente. Además, ofrece dos planes especiales, A y B. Por un plan de tipo A, obtiene una ganancia adicional de 8 dólares y por un plan de tipo B, 13 dólares. Además, se sabe que el 60 % de los clientes que contratan planes especiales prefieren uno de tipo A. Si una semana, la agencia vendió 25 viajes a la amazonía, 20 de los cuales no fueron especiales, ¿cuál es la ganancia esperada?

Solución: La agencia vendió 25 planes: 20 normales y 5 especiales.

Sea X la variable «número de planes de tipo A vendidos», $X \sim \mathbf{Bin}(5, 0.6)$.

La utilidad U que le producen los viajes vendidos es:

$$U = \underbrace{25 \times 6}_{\text{25 viajes}} + \underbrace{8X}_{\text{Plan A}} + \underbrace{13(5 - X)}_{\text{Plan B}}.$$

Dado que $\mathbf{E}(X) = np = 5 \times 0.6 = 3$, la utilidad esperada es:

$$\begin{aligned} \mathbf{E}(U) &= 150 + \mathbf{E}(8X) + 13\mathbf{E}(5 - X) = 215 - 5\mathbf{E}(X) \\ &= 215 - 5 \times 3 = 200. \end{aligned}$$

4.4. Distribuciones geométrica y binomial negativa

Consideremos una secuencia de pruebas de Bernoulli, con probabilidad de éxito p, pero en lugar de contar el número de éxitos, nos interesa conocer el número de intentos hasta obtener el primer éxito. Una sucesión de pruebas de este tipo se dice que forman un *experimento geométrico.*

Una variable aleatoria discreta X que puede tomar un número infinito de valores 1, 2, ..., se dice que sigue una ley de distribución geométrica de parámetro p $(0 < p < 1)$, si la probabilidad de que X tome el valor k es

$$\Pr(X = k) = p(1 - p)^{k-1}, \quad k = 1, 2, \ldots$$

A esta variable aleatoria se la nota como $X \sim \mathcal{G}(p)$. Su esperanza y su varianza son iguales a

$$\begin{aligned} \mathbf{E}(X) &= \frac{1}{p}. \\ \mathrm{Var}(X) &= \frac{1-p}{p^2}. \end{aligned}$$

La distribución geométrica se aplica en investigaciones de mercado y en muestreo, para conocer cuántas compras se han de realizar en una promoción para obtener un premio.

Ejemplos

1. Si la probabilidad de que un estudiante pase una prueba de ingreso a una universidad es 0.25, ¿cuál es la probabilidad de que el estudiante pase la prueba en el cuarto intento?

   Solución: En nuestro caso $p = 0.25$ y el número de intentos es $k = 4$, por lo que

   $$\begin{aligned} \Pr(X = 4) &= p(1 - p)^{4-1} \\ &= 0.25(1 - 0.25)^3 = 0.105. \end{aligned}$$

2. En una promoción una marca de papas fritas incluye, en cada una de las fundas, una de las figuras de los tres chiflados. Si un comprador cree que hay igual número de figuras de cada uno de los personajes en la promoción, ¿cuántas fundas ha de esperar comprar para obtener las tres figuras?

   Solución: En la primera compra, siempre obtiene una figura que no se tenía, por lo tanto $\mathbf{E}(X_1) = 1$.

   Para la segunda compra se tiene una probabilidad de $p_2 = \dfrac{2}{3}$ de conseguir una figura nueva; así, el número de compras que se debe realizar para obtenerla es $\mathbf{E}(X_2) = \dfrac{1}{p_2} = \dfrac{3}{2}$.

Una vez que se tienen dos figuras, la probabilidad de encontrar la figura que falta es $p_3 = \frac{1}{3}$ y el número de compras que se espera realizar para obtenerla es $\mathbf{E}(X_3) = \frac{1}{p_3} = 3$.

El número total de compras que se espera realizar es

$$\mathbf{E}(X) = \mathbf{E}(X_1) + \mathbf{E}(X_2) + \mathbf{E}(X_3) = 1 + 1.5 + 3 = 5.5.$$

Así, se espera realizar al menos 6 compras del producto para obtener la colección completa. ◀

Ahora, generalicemos la idea de la ley geométrica y nos interesa el número de pruebas de Bernoulli necesarias hasta obtener exactamente r éxitos.

Una variable aleatoria discreta X que puede tomar un número infinito de valores r, $r+1$, $r+2$, ..., se dice que sigue una ley de distribución binomial negativa de parámetros (r, p) $(r \geq 1,\ 0 < p < 1)$, si la probabilidad de que X tome el valor k es

$$\Pr(X = k) = \mathbf{C}_{k-1}^{r-1} p^r (1-p)^{k-r}, \quad k = r, r+1, r+2, \ldots$$

El parámetro r es el número de éxitos que se desea obtener y p es la probabilidad de obtener un éxito.

A esta variable aleatoria se la nota como $X \sim \mathbf{BN}(r, p)$. Su esperanza y su varianza son iguales a

$$\begin{aligned} \mathbf{E}(X) &= \frac{r}{p}. \\ \operatorname{Var}(X) &= r\frac{1-p}{p^2}. \end{aligned}$$

A la ley de distribución binomial negativa también se le llama *distribución de Pascal* y tiene las mismas aplicaciones que la ley geométrica.

Ejemplo. Una máquina, que está dañada, envasa latas de conserva de una en una y de manera independiente. Se considera que el 5 % de lo envasado resulta defectuoso. Si la máquina se detiene apenas produce el tercer defectuoso:

a) ¿Cuál es el número de latas producidas hasta que se detiene la máquina?

b) ¿Cuál es la probabilidad de que la máquina se detenga en la novena lata producida?

c) ¿Cuál es la probabilidad de que se detenga sin producir ninguna lata buena?

Solución: Definimos la variable aleatoria X: «número de latas producidas hasta que hayan 3 defectuosas»; $X \sim \mathbf{BN}(3, 0.05)$.

a) Calculemos la esperanza de X:

$$\mathbf{E}(X) = \frac{r}{p} = \frac{3}{0.05} = 60.$$

Se esperaría producir 60 latas hasta que se detenga la máquina.

b) Calculemos $\Pr(X = 9)$:

$$\Pr(X = 9) = \mathbf{C}_{9-1}^{3-1}(0.05)^3(1-0.05)^{9-3} = 0.00257.$$

c) Que ninguna lata producida fue buena, significa que las 3 primeras latas fueron defectuosas; es decir, $k = 3$.

$$\Pr(X = 3) = \mathbf{C}_{3-1}^{3-1}(0.05)^3(1-0.05)^{3-3} = 0.000125.$$

4.5. Distribución de Poisson

Una variable aleatoria discreta X que puede tomar un número infinito de valores 0, 1, 2, ..., se dice que sigue una ley de Poisson de parámetro λ ($\lambda > 0$), si la probabilidad de que X tome el valor k es

$$\Pr(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots$$

A esta variable aleatoria se la nota como $X \sim \mathcal{P}(\lambda)$.

Su esperanza y su varianza son, respectivamente, iguales a

$$\begin{aligned} \mathbf{E}(X) &= \lambda, \\ \mathrm{Var}(X) &= \lambda. \end{aligned}$$

La distribución de Poisson se aplica a sucesos que se presentan en el tiempo o en el espacio, tales como número de accidentes de tráfico, número de llamadas telefónicas a una central, número de goles que marca un equipo en un partido, número de bacterias en una placa, entre otros.

El significado del parámetro λ es el promedio de aparecimiento del evento en n pruebas.

Esta ley de probabilidad es una buena aproximación a la binomial cuando n es relativamente grande ($n \geq 30$) y p pequeño ($p \leq 0.05$), poniendo $\lambda = np$.

Puesto que muchas aplicaciones de esta ley dependen del tiempo, es conveniente ponerla en la siguiente forma

$$\Pr(X = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}, \quad k = 0, 1, 2, \ldots,$$

que se la interpreta como la probabilidad de que sucedan exactamente k eventos en un intervalo de tiempo fijo de duración t.

Para la ley de distribución de Poisson también existe una fórmula de recurrencia para el cálculo de las probabilidades[3], dada por

$$\begin{aligned} p_0 &= e^{-\lambda} \\ p_k &= p_{k-1} \times \frac{\lambda}{k}, \quad k = 1, 2, \ldots \end{aligned}$$

Ejemplos

1. En la construcción de un edificio el número de accidentes es de tres por mes. Calcular la probabilidad de que en un mes: a) hayan dos accidentes; b) hayan menos de 2 accidentes; c) hayan más de 3 accidentes.

   Solución: Si X es la variable «número de accidentes en un mes», $X \sim \mathcal{P}(3)$.

   a) La probabilidad solicitada es:

$$\Pr(X = 2) = \frac{e^{-3}3^2}{2!} = 0.224.$$

[3] Drane, J. W. y otros (1994), op. cit.

b) La probabilidad $\Pr(X < 2)$ es:

$$\begin{aligned}\Pr(X < 2) &= \Pr(X = 0) + \Pr(X = 1)\\ &= \frac{e^{-3}3^0}{0!} + \frac{e^{-3}3^1}{1!} = 0.199.\end{aligned}$$

c) Esta probabilidad es más cómodo calcularla mediante el evento complementario:

$$\begin{aligned}\Pr(X > 3) &= 1 - \Pr(X \leq 3) = 1 - [\Pr(X = 0) + \Pr(X = 1) + \Pr(X = 2) + \Pr(X = 3)]\\ &= 1 - \left(\frac{e^{-3}3^0}{0!} + \frac{e^{-3}3^1}{1!} + \frac{e^{-3}3^2}{2!} + \frac{e^{-3}3^3}{3!}\right) = 0.352.\end{aligned}$$

2. El promedio de llamadas que pasan por una central telefónica en un minuto es igual a dos. Hallar la probabilidad de que en tres minutos se hagan: a) 4 llamadas; b) menos de 4 llamadas; c) al menos 4 llamadas.

Solución: En este caso es necesario utilizar la segunda forma de la ley de Poisson con $\lambda = 2$ y $t = 3$:

$$\Pr(X = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}.$$

a) La probabilidad de que en 3 minutos se hagan 4 llamadas es

$$\Pr(X = 4) = \frac{e^{-2 \cdot 3}(2 \cdot 3)^4}{4!} = \frac{e^{-6}6^4}{24} = 0.1339.$$

b) La probabilidad buscada es:

$$\begin{aligned}\Pr(X < 4) &= \Pr(X = 3) + \Pr(X = 2) + \Pr(X = 1) + \Pr(X = 0)\\ &= \frac{6^3e^{-6}}{3!} + \frac{6^2e^{-6}}{2!} + \frac{6^1e^{-6}}{1!} + \frac{6^0e^{-6}}{0!}\\ &= 0.1512.\end{aligned}$$

c) Los eventos «se hicieron menos de 4 llamadas» y «se hicieron al menos 4 llamadas» son complementarios; por eso, su probabilidad es:

$$\Pr(X \geq 4) = 1 - \Pr(X < 4) = 1 - 0.1512 = 0.8488.$$

3. Un libro se edita con un tiraje de 1000 ejemplares. La probabilidad de que un libro esté encuadernado incorrectamente es igual a 0.01. Hallar la probabilidad de que el tiraje contenga exactamente cinco libros defectuosos.

Solución: Según los datos del problema $n = 1000$, $p = 0.01$ y $k = 5$. El número n es grande y p pequeño, por lo que utilizaremos la distribución de Poisson. Estimamos $\lambda = np = 1000 \times 0.01 = 10$.

La probabilidad buscada es

$$\Pr(X = 5) = \frac{e^{-10}10^5}{5!} = \frac{0.000045 \cdot 10^5}{120} = 0.0375.$$

4. El gerente de una fábrica planea comprar una máquina nueva de entre dos tipos A y B. Por cada día de funcionamiento, el número de reparaciones X que necesita la máquina A es una variable aleatoria de Poisson cuya media es $0.1t$, siendo t el tiempo de funcionamiento diario (en horas). El número de reparaciones diarias Y de la máquina B es una variable aleatoria de Poisson con media $0.12t$. El costo diario de operación de A es $C_A(t) = 10t + 30X^2$ y para B

es $C_B(t) = 8t + 30Y^2$. ¿Cuál de las máquinas da el menor costo esperado, si un día de trabajo consiste en: a) 10 horas? b) 20 horas?

Solución: El costo esperado para A es

$$\begin{aligned} \mathbf{E}(C_A(t)) &= \mathbf{E}(10t + 30X^2) = 10t + 30\mathbf{E}\left(X^2\right) \\ &= 10t + 30\left[\mathrm{Var}(X) + (\mathbf{E}(X))^2\right] \\ &= 10t + 30\left[0.1t + (0.1t)^2\right] \\ &= 13t + 0.3t^2. \end{aligned}$$

Igualmente,

$$\begin{aligned} \mathbf{E}(C_B(t)) &= \mathbf{E}(8t + 30Y^2) = 8t + 30\mathbf{E}\left(Y^2\right) \\ &= 8t + 30\left[\mathrm{Var}(Y) + (\mathbf{E}(Y))^2\right] \\ &= 8t + 30\left[0.12t + (0.12t)^2\right] \\ &= 11.6t + 0.432t^2. \end{aligned}$$

a) Calculemos $\mathbf{E}(C_A(10))$ y $\mathbf{E}(C_B(10))$:

$$\begin{aligned} \mathbf{E}(C_A(10)) &= 13(10) + 0.3(10)^2 = 160 \\ \mathbf{E}(C_B(10)) &= 11.6(10) + 0.432(10)^2 = 159.2. \end{aligned}$$

Si $t = 10$, el menor costo tiene la máquina B.

b) Calculemos $\mathbf{E}(C_A(20))$ y $\mathbf{E}(C_B(20))$:

$$\begin{aligned} \mathbf{E}(C_A(20)) &= 13(20) + 0.3(20)^2 = 380 \\ \mathbf{E}(C_B(20)) &= 11.6(20) + 0.432(20)^2 = 404.8. \end{aligned}$$

Si $t = 20$, la máquina A tiene una operación más económica.

4.6. Ejercicios

Ley uniforme discreta

1. Un reloj automático registra la hora a la cual llegan los empleados de una oficina, en horas y minutos completos. Una persona puede atrasarse hasta 59 minutos luego de la hora prefijada para entrar, caso contrario se le considera como falta. Por cada minuto de atraso se le cobra una multa de 50 centavos. Si los tiempos de atraso se consideran aleatorios:

    a) ¿Cuánto esperará una persona que se le descuente por un día que se atrasó?;

    b) Si en la oficina hay 8 personas, que se atrasaron 2 veces al mes cada una, ¿cuánto será el descuento global esperado a estos empleados de la oficina?

2. Para el servicio de transporte entre dos ciudades hay 10 buses, de los cuales 5 son de tipo normal (costo del pasaje 2 dólares) y 5 son de tipo especial (costo del pasaje 3 dólares). Una persona tiene que viajar entre las dos ciudades (ida y vuelta) durante los 5 días laborables de la semana, y para transportarse toma el primer bus que aparece en esa ruta, sin diferenciar el tipo; ¿cuánto esperará gastar esta persona en la semana?

3. En una escuela primaria se registró el número de palabras por minuto que leían los estudiantes, encontrándose que leían un mínimo de 80 palabras y un máximo de 139. Bajo la suposición de que la variable aleatoria que describe el número de palabras leídas está uniformemente distribuida.

   a) Halle la probabilidad de que un estudiante, seleccionado al azar, lea al menos 100 palabras;
   b) Determine el número de palabras que se esperaría lea un estudiante seleccionado al azar.

4. Sea X una variable aleatoria que sigue una ley uniforme sobre $\{-1, 0, 1\}$. Calcule: a) $\mathbf{E}\left(X^k\right)$ para $k = 1, 2, \ldots$; b) $\mathrm{Var}\left(X^k\right)$.

Ley hipergeométrica

5. Una variable aleatoria X tiene distribución hipergeométrica $\mathcal{H}(7, 4, 5)$. Calcule:

   a) $\Pr(X = 3)$;
   b) la esperanza utilizando la definición y verifíquela empleando la fórmula de $\mathbf{E}(X)$;
   c) la varianza de X.

6. En una línea de control de calidad se revisan 10 artículos, determinándose que hay 3 que no cumplen con las especificaciones. Si se escogen al azar dos artículos, identifique los parámetros de la ley y halle la esperanza de la variable aleatoria X, que describe el número de piezas correctas entre las dos escogidas.

7. Para llenar 4 vacantes de contador se presentan 10 personas, 7 hombres y 3 mujeres. Salen seleccionados 3 hombres y 1 mujer. Las mujeres acusan al empleador de discriminación sexual, por lo que le llevan a juicio. Si el juez supone que la elección fue al azar, ¿puede decirse que existió discriminación al hacer la elección?

8. El examen de graduación de los abogados consta de 50 temas. La forma de examinar es la siguiente: por sorteo se eligen 6 temas de los que hay que contestar 3 para aprobar. Si el estudiante ha estudiado solo 30 temas:

   a) ¿Cuál es la probabilidad de los 6 temas sepa contestar correctamente a 3?;
   b) ¿Cuál es la probabilidad de aprobar el examen?

9. Un auditor comprueba la contabilidad de una empresa y toma como muestra 3 cuentas de una lista de 8 cuentas por cobrar. Calcule la probabilidad de que el auditor encuentre por lo menos una cuenta vencida, si hay:

   a) 2 cuentas vencidas entre las 8 seleccionadas;
   b) 4 cuentas vencidas;
   c) 7 cuentas vencidas.

10. Una empresa renta autos, a los que no les da el mantenimiento debido, por lo que algunos funcionan mal. Un día tiene disponibles 8 autos para ser rentados, de los cuales 3 funcionan mal. Ese día se rentaron 4 autos. Calcule la probabilidad de que:

    a) ningún cliente haya recibido un auto que funcione mal;
    b) por lo menos un cliente reciba un auto que funcione mal;
    c) tres clientes reciban autos que funcionen mal.

Leyes de Bernoulli y binomial

11. Una variable aleatoria X tiene distribución binomial $\mathbf{Bin}(4, 0.2)$. Calcule:

a) $\Pr(X = 2)$;
b) $\Pr(X \geq 2)$;
c) $\Pr(X \leq 2)$;
d) $\mathbf{E}(X)$;
e) $\mathrm{Var}(X)$.

12. Una máquina llena las cajas de palillos de fósforo. En una proporción del 10 % la máquina no llena las cajas por completo. Se toman al azar 25 cajas de fósforos, calcule la probabilidad de que no haya más de dos cajas incompletas.

13. Una encuesta revela que el 20 % de la población es favorable a un político y el resto es desfavorable. Si se eligen seis personas al azar, se desea saber:

    a) La probabilidad de que las seis personas sean desfavorables;
    b) La probabilidad de que cuatro de las seis personas sean favorables.

14. Una determinada raza de perros tiene cuatro cachorros en cada camada. Si la probabilidad de que un cachorro sea macho es de 0.55, se pide calcular:

    a) la probabilidad de que en una camada dos exactamente sean hembras;
    b) la probabilidad de que en una camada al menos dos sean hembras.

15. En una investigación sobre el «rating» de los programas de televisión se detectó que se veía una telenovela en el canal 6 en un 28 % de los hogares. En una muestra aleatoria de diez hogares, halle la probabilidad de que:

    a) en cinco hogares se vea la telenovela del canal 6;
    b) ningún hogar esté sintonizando la novela.
    c) al menos en dos hogares se vea la novela.

16. En una instalación militar que dispone de 5 radares, la probabilidad de que un solo radar descubra a un avión de combate es de 0.7.

    a) ¿Cuál es la probabilidad de que sean exactamente 4 radares los que descubren al avión?;
    b) ¿Cuál es la probabilidad de que por lo menos uno lo descubra?;
    c) ¿De cuántos radares ha de constar la instalación para asegurarse en detectar aviones al menos en un 98 % de las veces?

17. Una aeronave dispone de 4 motores que funcionan independientemente, la probabilidad de que falle un motor durante el vuelo es 0.01. ¿Cuál es la probabilidad de que en un vuelo dado:

    a) no se observen fallas?;
    b) no se observe más de una falla?
    c) Si un avión puede seguir volando si al menos 2 motores continúan funcionando, ¿cuál es la probabilidad de que el avión se accidente?

18. Supóngase que la tasa de infección de una enfermedad contagiosa es del 25 %. En una oficina hay 10 personas que se vacunaron contra la enfermedad y ninguna se contagió.

    a) Determine la probabilidad de que ninguna persona se hubiera contagiado a pesar de que no se hubiera vacunado;
    b) De este resultado, ¿deduce usted que la vacuna es efectiva?

19. La probabilidad de que un estudiante obtenga el título de arquitecto es de 0.3. Calcule la probabilidad de que de un grupo de siete estudiantes matriculados en primer curso:

a) los siete finalicen la carrera;
b) al menos dos acaben la carrera;
c) ¿De cuántos alumnos ha de constar una promoción para asegurarse de que al menos uno culmine su carrera, con una probabilidad del 99 %?

20. En un estudio medioambiental se determinó que la presencia de mercurio en el agua envenena al 20 % de los peces en 24 horas. Para confirmar el resultado se colocaron 20 peces en un tanque con agua contaminada. Calcule la probabilidad de que en 24 horas:

 a) sobrevivan exactamente 14 peces;
 b) sobrevivan por lo menos 10 peces
 c) sobrevivan cuando mucho 16 peces;
 d) Calcule el número de peces que se espera que sobrevivan;
 e) Calcule la varianza del número de sobrevivientes.

21. Una compañía petrolera va a perforar 29 pozos, cada uno de ellos tiene una probabilidad de 0.1 de producir petróleo de manera rentable. A la compañía le cuesta 100 mil dólares perforar cada pozo. Un pozo comercial extrae petróleo por un valor de 5 millones de dólares. Calcule:

 a) la ganancia que espera obtener la compañía por los 29 pozos;
 b) la desviación estándar del valor de la ganancia.

22. Una línea aérea, habiendo observado que el 5 % de las personas que hacen reservación no se presentan para el vuelo, vende 100 boletos para un avión que tiene 95 asientos. ¿Cuál es la probabilidad de que, el momento del vuelo, haya un asiento disponible para cada pasajero?

23. En un examen se plantean 10 preguntas a las que debe responderse verdadero o falso. Un alumno aprobará el examen si al menos 7 respuestas son acertadas. ¿Qué probabilidad de aprobar tiene un estudiante que responde todo al azar? ¿Y uno que sabe el 30 % de la asignatura?

Leyes geométrica y binomial negativa

24. Cuando se graba un comercial de televisión, la probabilidad de que un actor recite correctamente el diálogo de su toma es 0.3. ¿Cuál es la probabilidad que el actor recite correctamente su diálogo en la sexta vez?

25. En un examen el profesor realiza varias preguntas a un estudiante. La probabilidad de que el estudiante responda correctamente a cualquier pregunta es igual a 0.9. El profesor interrumpe el examen apenas el estudiante manifiesta el desconocimiento de la pregunta hecha. Se requiere:

 a) formar la ley de distribución de la variable aleatoria que describe el número de preguntas que realiza el profesor;
 b) hallar el número esperado de preguntas que ha de realizar el profesor.

26. La probabilidad de que un tirador haga blanco en un solo disparo es igual a 0.2. Al tirador se le entregan cartuchos hasta tanto no yerre el tiro.

 a) Forme la ley de distribución que describe el número de cartuchos utilizados;
 b) ¿Cuántos cartuchos se espera que utilice el tirador?

27. En un examen, en el que se realizan preguntas sucesivas, para aprobar hay que contestar correctamente a 10 preguntas. Suponiendo que el alumno sepa el 80 % de las respuestas, ¿cuál es la probabilidad de que apruebe en las 12 primeras preguntas?

28. Una jugadora de tenis gana el 33 % de los partidos que realiza. Ella jugará en un torneo mientras no sea eliminada por perder un partido.

    a) Halle la probabilidad de que sea eliminada en el segundo partido;
    b) Si para ganar el torneo se deben ganar 5 partidos consecutivos, ¿cuál es la probabilidad de que la jugadora pierda en la final del torneo?;
    c) ¿Cuántos partidos se espera que llegue a jugar durante el torneo?

29. Una marca de refrescos tiene impresas, en cada una de las tapas, una de las figuras de los 4 jinetes del apocalipsis, y quien reuna la colección completa ganará un premio. Si un comprador cree que hay igual número de figuras de cada uno de los personajes en la promoción, ¿cuántos refrescos ha de esperar comprar para ganar el premio?

30. Un pájaro de cierta especie come gusanos de una población muy grande. Estos gusanos pueden comer, a su vez, de una planta venenosa, de manera que si el pájaro come un gusano envenenado, deja de comer gusanos ese día. Suponiendo que el 33 % de la población de gusanos come de la planta venenosa, hallar el número medio de gusanos comidos por un pájaro en un día.

31. Un lepidopterista solo está interesado en los ejemplares de una clase de mariposas, que constituyen el 15 % de todas las mariposas de la zona. Halle la probabilidad de que esta persona tenga que cazar 8 mariposas de las que no le interesan antes de encontrar:

    a) un ejemplar de la clase deseada;
    b) tres ejemplares de la clase deseada.

32. En una fábrica, el departamento de control de calidad, revisa los lotes de piezas que entran, de acuerdo con el siguiente criterio: se van extrayendo piezas sucesivamente y el lote es rechazado si se encuentra la primer pieza defectuosa antes de la vigésima extracción. Si conocemos que el 2 % de piezas son defectuosas, ¿cuál es la probabilidad de que un lote sea rechazado?

33. En una fábrica, se examinan las piezas que salen de una determinada máquina. Supongamos que si en una hora salen mas de 5 piezas defectuosas, la máquina debe ser recalibrada. Si suponemos que la probabilidad de que una pieza sea defectuosa es 0.2, y es la misma para todas las piezas fabricadas; encontrar:

    a) la probabilidad de que se tenga que recalibrar la máquina cuando se han inspeccionado 20 piezas;
    b) la probabilidad de que se recalibre la máquina sin haber producido ninguna pieza buena;
    c) El número esperado de piezas que se deben inspeccionar.

34. Se sabe que, aproximadamente, el 20 % de los usuarios de Windows no cierran el programa adecuadamente. Supongamos que el Windows está instalado en una computadora pública que es utilizada aleatoriamente por personas que actúan independientemente unas de otras.

    a) ¿Cuál es la probabilidad de que el tercer usuario sea el primero que cierra adecuadamente el Windows?;
    b) ¿Cuál es el número medio de personas que usan la computadora desde el momento en que se enciende hasta que alguien no cierra el programa adecuadamente?

Ley de Poisson

35. Sea Y una variable aleatoria que sigue una distribución de Poisson de media $\lambda = 2$. Calcule:

a) $\Pr(Y = 4)$;
b) $\Pr(Y \leq 4)$;
c) $\Pr(Y > 4)$;
d) $\Pr(Y \geq 4 | Y \geq 2)$.

36. El promedio de llamadas que recibe una central telefónica en un minuto es de 1.5. Halle la probabilidad de que en cuatro minutos se reciban:

a) 3 llamadas;
b) menos de 3 llamadas;
c) no menos de cuatro y no más de siete.

37. Suponga que el número de pacientes que ingresan a la sala de emergencia de un hospital en la noche del viernes tiene una distribución de Poisson con media igual a 4. Evalúe las probabilidades de que:

a) durante una noche haya exactamente 2 pacientes en la sala de emergencia;
b) durante la noche hayan más de 3 personas.

38. En un hotel, el promedio de pedidos de servicio a la habitación es igual a 2 cada media hora. Halle la probabilidad de que en una hora se reciban:

a) 3 pedidos;
b) menos de 3 pedidos;
c) no menos de 3 pedidos.

39. Una fábrica de gaseosas recibió 100 botellas vacías. La probabilidad de que al transportarlas resulte una botella rota es 0.03. Halle la probabilidad de que la fábrica reciba rotas:

a) exactamente dos botellas;
b) más de dos;
c) por lo menos una.

40. El promedio de automóviles que entran en un túnel en una montaña es de un carro cada 2 minutos. Si un número excesivo de autos entra al túnel en un período corto de tiempo se genera una situación peligrosa. Encuentre la probabilidad de que el número de autos que entran al túnel en un período de 2 minutos exceda de tres.

41. Se supone que el número de bacterias por mm^3 de agua en un estanque es una variable aleatoria X con distribución de Poisson de parámetro $\lambda = 0.5$.

a) ¿Cuál es la probabilidad de que en $1\,\text{mm}^3$ de agua del estanque no haya ninguna bacteria?;
b) En 40 tubos de ensayo se toman muestras de agua del estanque ($1\,\text{mm}^3$ de agua en cada tubo). ¿Qué distribución sigue la variable Y: «número de tubos de ensayo, entre los 40, que no contienen bacterias»? Calcule $\Pr(Y \geq 20)$;
c) Si sabemos que en un tubo hay bacterias, ¿cuál es la probabilidad de que haya menos de tres?

42. Una planta embotelladora de refrescos tiene una máquina vieja para llenar botellas. La máquina produce una ganancia de 100 dólares por día de trabajo; sin embargo, se descompone en promedio 2 veces cada 10 días. Si Y representa el número de descomposturas durante el funcionamiento de la máquina y t es el número de días que trabajó la máquina, la ganancia generada por la máquina se expresa por $G = 100t - 50Y^2$. Determine la ganancia esperada en 10 días de trabajo.

43. En una población el 1 por ciento de la población sufre de daltonismo, ¿cuál es la probabilidad de que entre 100 personas:

a) ninguna padezca de daltonismo;
b) 2 o más lo padezcan;

c) ¿Cuán grande debe ser una muestra aleatoria (con reemplazo) para que la probabilidad de que al menos una persona tenga daltonismo sea mayor o igual a 0.95?

44. En un bosque de cedro el número de árboles con plaga por hectárea Y tiene una distribución de Poisson $\mathcal{P}(10)$. Los árboles con plaga se tratan con insecticida a un costo de 3 dólares por árbol; además, de un costo fijo, por uso del equipo y transporte, igual a 50 dólares. Halle el valor esperado y la desviación estándar del costo total C de fumigar 5 hectáreas de bosque.

45. Para el control de calidad de discos para computadora se emplea un dispositivo electrónico que cuenta el número de bytes defectuosos. Una marca de discos de computadora tiene un promedio de 0.1 bytes defectuosos por disco. Calcule el porcentaje de discos que:

   a) no tienen defectos;
   b) tienen algún defecto;
   c) Halle la probabilidad de que ninguno de dos discos inspeccionados, ninguno tenga defectos.

46. El número de automóviles que llegan a un estacionamiento, que tiene una capacidad de 12 autos, es una variable aleatoria que sigue una ley de Poisson, con un promedio de 4 por hora. Si al inicio del día el estacionamiento está vacío,

   a) ¿cuál es la probabilidad de que se llene durante la primera hora?;
   b) Calcule la probabilidad de que lleguen menos de 30 vehículos en un turno de 8 horas.

47. Si hay en promedio, un 1 por ciento de zurdos, ¿cuál es la probabilidad de tener por lo menos 4 zurdos entre 200 personas?

48. En una investigación de mercado se determinó que el 2 por ciento de la población toma regularmente una marca de yogurt. Se escogió una muestra de 300 personas, determine la probabilidad de que:

   a) exactamente 5 personas tomen yogurt de esa marca;
   b) a lo más tomen 3 personas;
   c) al menos tomen 5 personas.

49. La tasa mensual de suicidios es de 4 por un millón de personas. En una ciudad de 500 000 habitantes, halle probabilidad de que:

   a) en un mes dado, hayan menos de 5 suicidios;
   b) ¿Será sorprendente que durante un año, al menos en dos meses ocurran más de 4 suicidios?

50. En estudios demográficos sobre matrimonios que tienen algún tipo de planificación familiar, el número X de hijos por matrimonio es igual a 2, salvo ciertas desviaciones debidas al azar. Se ha comprobado que, o bien

$$X = 2 - (Y + 1),$$

donde Y es una variable de Bernoulli de parámetro $p = 0.3$, y ésto ocurre con probabilidad $p = \frac{1}{2}$ (pues se cumple en el 50 % de los matrimonios), o bien es

$$X = 2 + Z,$$

donde Z sigue una distribución de Poisson de parámetro λ; y esto segundo ocurre con también con probabilidad $p = \frac{1}{2}$. Halle:

   a) el valor de λ, sabiendo que $\mathbf{E}(X) = 2$;
   b) la probabilidad de que un matrimonio tenga uno o dos hijos.

4.7. Distribución uniforme

La ley de distribución de probabilidad de una variable aleatoria continua X se llama *uniforme* si en el intervalo $[a, b]$ la función de densidad es constante e igual a

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & \text{si } x \in [a,b] \\ 0, & \text{si } x \notin [a,b] \end{cases}$$

A esta variable aleatoria se la nota como $X \sim \mathcal{U}[a, b]$.

La función de distribución es

$$F(x) = \begin{cases} 0, & \text{si } x < a; \\ \dfrac{x-a}{b-a}, & \text{si } a \leq x \leq b; \\ 1, & \text{si } x > b. \end{cases}$$

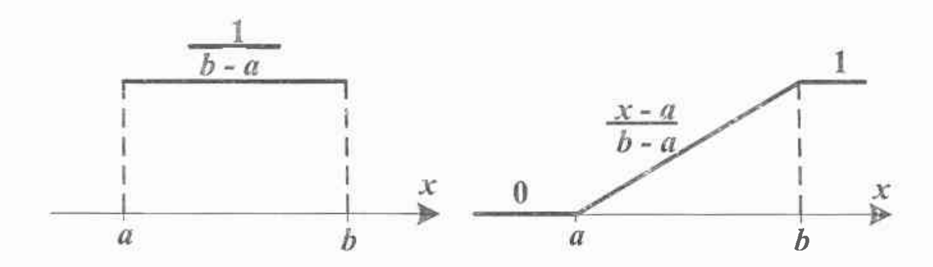


Figura 4.1: Funciones de densidad y de distribución de una variable aleatoria uniforme.

La esperanza y la varianza son iguales a:

$$\begin{aligned} \mathrm{E}(X) &= \frac{a+b}{2}, \\ \mathrm{Var}(X) &= \frac{(b-a)^2}{12}. \end{aligned}$$

Esta ley es el análogo continuo de la distribución uniforme discreta, que asigna igual probabilidad a cada resultado de un experimento. Tiene amplia aplicación en problemas de simulación estadística y en fenómenos que presentan regularidad en su aparecimiento, pero que no es posible usar variables discretas, como cuando dependen del tiempo. También, el error originado por el redondeo de un número se describe satisfactoriamente mediante una ley uniforme en el intervalo $\left[-\dfrac{1}{2}, \dfrac{1}{2}\right]$.

Ejemplos

1. Una variable aleatoria X sigue una distribución uniforme sobre $[-2, 3]$.

   a) Calcular $\Pr(X = 1)$, $\Pr(X < 1.3)$, $\Pr(|X| < 1.5)$;

   b) Hallar un valor de t tal que $\Pr(X > t) = \dfrac{1}{3}$.

   Solución:

a) Estimemos la función de densidad.
Sabemos que $f(x) = 0$ si $x \notin [-2, 3]$.
Se tienen los límites $a = -2$, $b = 3$; por lo que $f(x) = \dfrac{1}{3-(-2)} = \dfrac{1}{5}$, en $[-2, 3]$.
La función de densidad queda así:

$$f(x) = \begin{cases} \dfrac{1}{5}, & \text{si } x \in [-2, 3]; \\ 0, & \text{si } x \notin [-2, 3]. \end{cases}$$

- $\Pr(X = 1) = \displaystyle\int_1^1 f(x)\,dx = 0$, porque X es una variable aleatoria continua.
- $\Pr(X < 1.3) = \displaystyle\int_{-\infty}^{1.3} f(x)\,dx = \int_{-2}^{1.3} \frac{1}{5}\,dx = 0.66.$
- $\Pr(|X| < 1.5) = \Pr(-1.5 < X < 1.5) = \displaystyle\int_{-1.5}^{1.5} f(x)\,dx = \int_{-1.5}^{1.5} \frac{1}{5}\,dx = \frac{3}{5} = 0.6.$

b) Calculemos $\Pr(X > t)$:

$$\begin{aligned} \Pr(X > t) &= \int_t^{\infty} f(x)\,dx = \int_t^3 \frac{1}{5}\,dx + \int_3^{\infty} 0\,dx \\ &= \left[\frac{x}{5}\right]_t^3 = \frac{3-t}{5}. \end{aligned}$$

Entonces, $\dfrac{3-t}{5} = \dfrac{1}{3}$, con lo cual $t = \dfrac{4}{3}$.

2. Dos amigos, Roberto y Fernando, deben encontrarse en una parada de bus entre las 9:00 y 10:00 h. Cada uno esperará un máximo de 10 minutos. ¿Cuál es la probabilidad de que no se encuentren, si Fernando llegará a las 9:30 en punto?

Solución: La variable aleatoria X que describe el tiempo de llegada de Roberto puede tomar cualquier valor entre las 9:00 y 10:00 h o entre 0 y 60 minutos. De manera que $X \sim \mathcal{U}[a, b]$ y su función de densidad es

$$f(t) = \begin{cases} \dfrac{1}{60}, & \text{si } 0 \le t \le 60; \\ 0, & \text{caso contrario.} \end{cases}$$

Puesto que Fernando llegará a las 9:30 o a los 30 minutos después de las 9 y esperará a lo más 10 minutos, Roberto no se encontrará con Fernando si llega de 9:00 a menos de 9:20 o si llega después de las 9:40.

Entonces, la probabilidad de que no se encuentren es

$$\begin{aligned} \Pr\left[(0 \le X < 20) \text{ o } (40 < X \le 60)\right] &= \int_0^{20} \frac{1}{60}\,dt + \int_{40}^{60} \frac{1}{60}\,dt \\ &= \frac{1}{3} + \frac{1}{3} = 0.667. \end{aligned}$$

3. En una empresa farmacéutica un proceso se detiene cuando deja de funcionar un esterilizador hasta que llegue su repuesto. El tiempo de entrega T está uniformemente distribuido en un intervalo de uno a cinco días. El costo C de esa falla y la parada comprende un costo fijo de 200 dólares de la refacción y un costo variable que aumenta en proporción a T^2, de modo que

$$C = 200 + 12T^2.$$

a) Calcular la probabilidad de que el tiempo de espera sea de dos días o más.

b) Calcular el costo esperado de una falla.

Solución:

a) El tiempo de entrega está uniformemente distribuido de uno a cinco días, de modo que $a = 1$ y $b = 5$:

$$f(t) = \begin{cases} \frac{1}{4}, & \text{si } 1 \leq t \leq 5; \\ 0, & \text{caso contrario.} \end{cases}$$

Así,

$$\Pr(T \geq 2) = \int_2^5 f(t)\,dt = \int_2^5 \frac{1}{4}\,dt = \frac{1}{4}(5-2) = \frac{3}{4}.$$

b) Por las propiedades de la esperanza, $\mathbf{E}(C) = 200 + 12\mathbf{E}(T^2)$. Calculemos $\mathbf{E}(T^2) = \text{Var}(T) + (\mathbf{E}(T))^2$:

$$\begin{aligned} \mathbf{E}(T^2) &= \frac{(b-a)^2}{12} + \left(\frac{a+b}{2}\right)^2 \\ &= \frac{(5-1)^2}{12} + \left(\frac{1+5}{2}\right)^2 = \frac{31}{3}. \end{aligned}$$

Así,

$$\mathbf{E}(C) = 200 + 12\left(\frac{31}{3}\right) = 200 + 124 = 324.$$

El costo esperado de una falla es de 324 dólares.

4.8. Distribución exponencial

Se dice que una variable aleatoria continua X sigue una ley de distribución *exponencial* de parámetro λ si su función de densidad es

$$f(x) = \begin{cases} 0, & \text{si } x < 0; \\ \lambda e^{-\lambda x}, & \text{si } x \geq 0; \end{cases}$$

donde λ es una constante positiva. Notaremos como $X \sim \mathcal{E}(\lambda)$.

La función de distribución correspondiente es

$$F(x) = \begin{cases} 0, & \text{si } x < 0 \\ 1 - e^{-\lambda x}, & \text{si } x \geq 0 \end{cases}$$

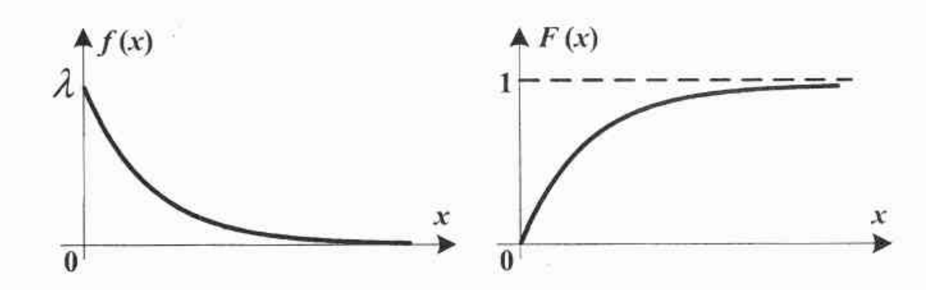


Figura 4.2: Funciones de densidad y de distribución de una variable aleatoria exponencial.

La esperanza y la varianza son iguales a:

$$\mathbf{E}(X) = \frac{1}{\lambda}, \qquad \operatorname{Var}(X) = \frac{1}{\lambda^2}.$$

Esta ley surge en problemas de genética, duración de aparatos electrónicos o desintegración radiactiva. También, es la principal en la teoría de los procesos de Markov.

Relación entre las distribuciones exponencial y de Poisson

Sea X la variable que cuenta el número de eventos que ocurren en el tiempo $[0, t]$, con media λt: entonces,

$$\Pr(X = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}, \quad k = 0, 1, 2, \ldots$$

Sea T el tiempo que transcurre hasta que sucede el primer evento de Poisson. El rango de T es el intervalo $[0, \infty[$ y su función de distribución es

$$F_T(t) = \Pr(T \leq t) = 1 - \Pr(T > t) = 1 - \Pr(X = 0) = 1 - e^{-\lambda t},$$

donde el evento $(T > t)$ indica que el primer evento de Poisson ocurre después de t, o lo que es lo mismo, que no ocurre ningún evento en el intervalo $[0, t]$; es decir, $(T > t) = (X = 0)$.

También, se tiene que

$$f_T(t) = F_T'(t) = \lambda e^{-\lambda t}, \quad t > 0.$$

Ejemplos

1. Una variable aleatoria continua Y está distribuida según una ley exponencial $\mathcal{E}(3)$.

   a) Hallar la probabilidad del resultado de la prueba: Y esté en el intervalo $(0.13; 0.7)$;

   b) Determinar su esperanza y su desviación estándar.

   Solución: Como $Y \sim \mathcal{E}(3)$,

   $$f(y) = \begin{cases} 0, & \text{si } y < 0; \\ 3e^{-3y}, & \text{si } y \geq 0. \end{cases}$$

   a) Determinemos $\Pr(0.13 < Y < 0.7)$:

   $$\begin{aligned} \Pr(0.13 < Y < 0.7) &= \int_{0.13}^{0.7} f(y)\,dy = \int_{0.13}^{0.7} 3e^{-3y}\,dy \\ &= 3\left(-\frac{1}{3}\right) e^{-3y}\Big|_{0.13}^{0.7} \\ &= 0.677 - 0.122 = 0.552. \end{aligned}$$

   b) Se tiene que

   $$\begin{aligned} \mathbf{E}(Y) &= \frac{1}{\lambda} = \frac{1}{3}, \\ \sigma(Y) &= \sqrt{\operatorname{Var}(Y)} = \sqrt{\frac{1}{9}} = \frac{1}{3}. \end{aligned}$$

2. El tiempo durante el cual las baterías para teléfono celular trabajan en forma efectiva hasta que fallan se distribuye según un modelo exponencial, con un tiempo promedio de falla de 500 horas.

   a) Calcular la probabilidad de que una batería funcione por más de 600 horas;

   b) Si una batería ha trabajado 350 horas, ¿cuál es la probabilidad de que trabaje más de 300 horas adicionales?

Solución: Consideremos la variable aleatoria X: «tiempo que dura la batería hasta que falla». Como $\mathbf{E}(X) = 500 = \frac{1}{\lambda}$, entonces $\lambda = \frac{1}{500}$ y $X \sim \mathcal{E}\left(\frac{1}{500}\right)$. Su función de distribución es:

$$F(x) = \begin{cases} 0, & \text{si } x < 0; \\ 1 - e^{-x/500}, & \text{si } x \geq 0. \end{cases}$$

a) Calculemos $\Pr(X > 600)$:

$$\begin{aligned} \Pr(X > 600) &= 1 - \Pr(X \leq 600) = 1 - F(600) \\ &= 1 - \left(1 - e^{-600/500}\right) = 0.301. \end{aligned}$$

b) Si la batería ya trabajó 350 horas, queremos conocer la probabilidad de que trabaje más de 650:

$$\begin{aligned} \Pr(X > 350 + 300 | X > 350) &= \frac{\Pr(X > 650)}{\Pr(X > 350)} = \frac{1 - F(650)}{1 - F(350)} \\ &= \frac{1 - \left(1 - e^{-650/500}\right)}{1 - \left(1 - e^{-350/500}\right)} = 0.549. \end{aligned}$$

3. El número de clientes que llega a una ventanilla de un banco sigue una distribución de Poisson con media de 4 personas por minuto. ¿Cuál es la probabilidad de que el primer cliente llegue:

   a) dentro de los 30 segundos después de haber abierto la ventanilla?

   b) más de dos minutos después de abrir la ventanilla?

Solución: Tenemos las siguientes variables aleatorias:

X: «número de clientes que llegan a la ventanilla», $X \sim \mathcal{P}(4)$.

T: «tiempo que transcurre hasta que llega el primer cliente», $T \sim \mathcal{E}(4)$, con función de densidad

$$f_T(t) = 4e^{-4t}, \quad t > 0.$$

a) La probabilidad de que el primer cliente llegue en los primeros 30 segundos es

$$\Pr(T \leq 0.5) = \int_0^{0.5} 4e^{-4t} dt = 0.865.$$

Obsérvese que también se cumple que

$$\Pr(T \leq 0.5) = 1 - \Pr(T > 0.5) = 1 - \Pr(T = 0) = 1 - e^{-4(0.5)} = 0.865.$$

b) Se tiene que

$$\Pr(T > 2) = \int_2^{\infty} 4e^{-4t} dt = 0.000335.$$

4.9. Distribución normal

La ley de probabilidad de una variable aleatoria continua X se llama *normal* si su función de densidad es

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \quad x \in (-\infty, \infty)$$

donde μ es un valor real cualquiera y σ es positivo. A tal variable aleatoria se notará como $X \sim \mathcal{N}(\mu, \sigma^2)$.

La función de distribución correspondiente es la integral

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-(t-\mu)^2/2\sigma^2}\, dt.$$

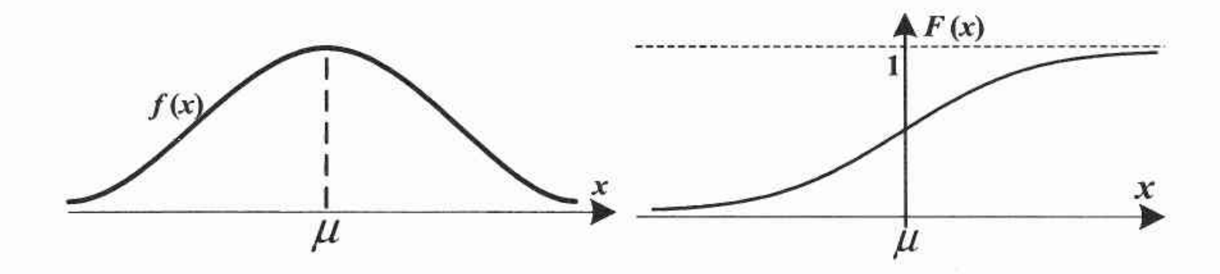


Figura 4.3: Funciones de densidad y de distribución de una variable aleatoria normal.

La esperanza y la varianza son iguales a

$$\mathbf{E}(X) = \mu, \qquad \mathrm{Var}(X) = \sigma^2.$$

Por esto, se dice que es una ley normal de media μ y varianza σ^2. Obviamente σ es la desviación estándar de X.

Observemos que la función de densidad de una variable aleatoria $X \sim \mathcal{N}(\mu, \sigma^2)$ es simétrica respecto a la recta $x = \mu$.

Esta ley tiene amplia aplicación en física, economía, ingeniería y biología, pues –como una primera aproximación– se asume que los fenómenos siguen una ley normal. También, juega un papel muy importante en toda la teoría estadística ya que, bajo amplias suposiciones, el comportamiento de las sumas de magnitudes aleatorias es aproximadamente normal, lo que constituye el Teorema del Límite Central.

El nombre de *normal* fue aplicado a esta ley de distribución por F. Galton en 1889, no sin reparos por otros científicos, porque este nombre puede hacer pensar a las personas que las otras distribuciones son, en uno u otro sentido, anormales. En el plano anecdótico, remitámonos a lo que se dice en el libro de Mosteller, Rourke y Thomas (1970, p. 226) respecto al nombre de esta ley: «Algunas veces la distribución normal es llamada gaussiana, especialmente en la ingeniería y la física. En Francia es llamada laplaciana. Estos nombres son usados, probablemente, porque la distribución fue inventada por de Moivre.»

Un caso importante de esta ley de probabilidad se tiene cuando $\mu = 0$ y $\sigma^2 = 1$, que se denomina

normal estándar ($\mathcal{N}(0,1)$), sus funciones de densidad y distribución son

$$\begin{aligned}\varphi(x) &= \frac{1}{\sqrt{2\pi}}e^{-x^2/2}, \qquad x \in (-\infty, \infty)\\ \Phi(x) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}\,dt,\end{aligned}$$

respectivamente. Obsérvese que, en este caso particular, la función de densidad se nota mediante φ y la función de distribución por Φ.

Si se tiene una variable aleatoria $X \sim \mathcal{N}(\mu, \sigma^2)$, pueden calcularse los valores de su función de distribución mediante el empleo de la ley normal estándar aplicando la transformación

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

La función de distribución de la ley normal estándar no se puede dar como una función explícita, sino solamente en forma de una integral, por lo que se emplean tablas, como la que se encuentra en la Tabla 1 del Apéndice, para calcular los valores de $\Phi(x)$.

Si $X \sim \mathcal{N}(\mu, \sigma^2)$, se puede dar la siguiente regla empírica que da el área bajo la curva limitada por una, dos y tres veces la desviación estándar (ver Figura 4.4).

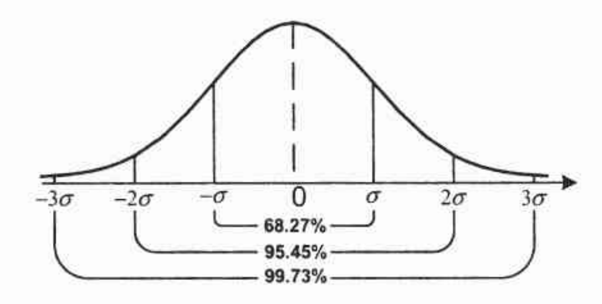


Figura 4.4: Áreas bajo la curva de la distribución normal.

1. El área limitada por el intervalo $[\mu-\sigma, \mu+\sigma]$ contiene un área igual a 0.682.
2. El área limitada por el intervalo $[\mu-2\sigma, \mu+2\sigma]$ contiene un área igual a 0.954.
3. El área limitada por el intervalo $[\mu-3\sigma, \mu+3\sigma]$ contiene un área igual a 0.997.

Ejemplos

1. La esperanza de una variable aleatoria X normal es igual a 6 y su varianza es 16. Escribir la ley de la variable aleatoria y calcular:

   a) $\Pr(X < 3)$; b) $\Pr(X > 4)$; c) $\Pr(4.5 < X < 7)$;

   d) Encontrar el valor de t de manera que se cumpla que $\Pr(X < t) = 0.9264$.

   Solución: La esperanza es $\mathbf{E}(X) = \mu = 6$ y la varianza es $\sigma^2 = 16$, por lo que $\sigma = 4$; entonces,

$$f(x) = \frac{1}{\sqrt{2\pi}\cdot 4}\exp\left(-\frac{(x-6)^2}{2\cdot 16}\right) = \frac{1}{4\sqrt{2\pi}}\exp\left(-\frac{(x-6)^2}{32}\right).$$

a) Calculemos $\Pr(X < 3)$:

$$\begin{aligned}\Pr(X < 3) &= F(3) = \Phi\left(\frac{3-6}{4}\right) \\ &= \Phi(-0.75) = 0.2266.\end{aligned}$$

b) Calculemos $\Pr(X > 4)$:

$$\begin{aligned}\Pr(X > 4) &= 1 - \Pr(X \leq 4) = 1 - F(4) \\ &= 1 - \Phi\left(\frac{4-6}{4}\right) = 1 - \Phi(-0.5) \\ &= 1 - 0.3085 = 0.6915.\end{aligned}$$

c) Calculemos $\Pr(4.5 < X < 7)$:

$$\begin{aligned}\Pr(4.5 < X < 7) &= F(7) - F(4.5) \\ &= \Phi\left(\frac{7-6}{4}\right) - \Phi\left(\frac{4.5-6}{4}\right) = \Phi(0.25) - \Phi(-0.375) \\ &= 0.5987 - 0.3538 = 0.2449.\end{aligned}$$

d) Se tiene que

$$\Pr(X < t) = \Phi\left(\frac{t-6}{4}\right) = 0.9264.$$

Por otro lado, en la tabla de la ley normal, se encuentra que $\Phi(1.45) = 0.9264$; es decir, se cumple que

$$\frac{t-6}{4} = 1.45.$$

Entonces, $t = 1.45 \times 4 + 6 = 11.8$.

2. El perímetro craneal de los hombres, en una ciudad, es una variable aleatoria de media 60 cm y desviación estándar 2 cm.

a) ¿Qué porcentaje de los hombres tienen un perímetro craneal entre 57 y 64 cm?;
b) ¿Qué perímetro craneal debe tener un hombre para que el 16.6 % de sus paisanos «tengan más cabeza que él»?;
c) ¿Y cuánto para que el 35.2 % tenga menos?

Solución: Se tiene la variable aleatoria X: «perímetro craneal de un individuo», $X \sim \mathcal{N}(60, 2^2)$.

a) Buscamos $\Pr(57 \leq X \leq 64)$:

$$\begin{aligned}\Pr(57 \leq X \leq 64) &= F(64) - F(57) \\ &= \Phi\left(\frac{64-60}{2}\right) - \Phi\left(\frac{57-60}{2}\right) = \Phi(2) - \Phi(-1.5) \\ &= 0.9772 - 0.0668 = 0.9104.\end{aligned}$$

El 91 % de la población tiene un perímetro craneal entre 57 y 64 cm.

b) Debemos hallar el valor de x de manera que $\Pr(X > x) = 0.166$.

$$\begin{aligned}\Pr(X > x) &= 1 - \Pr(X \leq x) = 1 - F(x) = 0.166, \\ F(x) &= 1 - 0.166 = 0.834, \\ \Phi\left(\frac{x-60}{2}\right) &= 0.834.\end{aligned}$$

Por otro lado, en la tabla de la ley normal estándar se observa que $\Phi(0.97) = 0.834$. Entonces,

$$\frac{x-60}{2} = 0.97.$$

De donde, $x = 61.94\,\text{cm}$.

c) Ahora, hallemos x tal que $\Pr(X < x) = 0.352$.

$$\begin{aligned} \Pr(X < x) &= F(x) = 0.352, \\ \Phi\left(\frac{x-60}{2}\right) &= 0.352. \end{aligned}$$

En la tabla leemos que $\Phi(-0.38) = 0.352$. Entonces,

$$\frac{x-60}{2} = -0.38.$$

Por tanto, $x = 59.24\,\text{cm}$.

3. **En una fábrica de autos un ingeniero está diseñando autobuses pequeños. Sabe que la estatura de la población está normalmente distribuida con media 1.70 m y varianza σ^2, con $\sigma = 5\,\text{cm}$. ¿Qué altura mínima deberán tener los autobuses para que no más del 1 % de las personas golpee su cabeza con la parte superior del autobús?**

Solución: Sea X la variable aleatoria «estatura de las personas», $X \sim \mathcal{N}(1.70, (0.05)^2)$. Denominemos h a la altura mínima para que la probabilidad de que una persona golpee su cabeza con el techo del autobús sea del 1 %; es decir,

$$\begin{aligned} \Pr(X > h) &= 0.01 \\ \Pr(X > h) &= 1 - \Pr(X \le h) = 1 - F(h) = 1 - \Phi\left(\frac{h-1.70}{0.05}\right), \end{aligned}$$

donde F es la función de distribución de la variable aleatoria X. Ahora bien,

$$\begin{aligned} 1 - \Phi\left(\frac{h-1.70}{0.05}\right) &= 0.01. \\ \Phi\left(\frac{h-1.70}{0.05}\right) &= 1 - 0.01 = 0.99. \end{aligned}$$

Igualando los argumentos de la función Φ da:

$$\begin{aligned} \frac{h-1.70}{0.05} &= 2.33. \\ h &= 1.70 + (0.05)(2.33) = 1.817. \end{aligned}$$

Es decir, el ingeniero deberá diseñar el autobús con una altura de 1.82 m.

4. **En una ciudad habitan 150 mil familias, cuyo ingreso anual sigue distribución normal con media de 8000 dólares y desviación estándar de 1200 dólares.**

   a) **¿Cuántas familias tendrán ingresos anuales menores a 6600 dólares?;**

   b) **El SRI dice que pagarán sus impuestos aquellas familias que se encuentren en el último quintil. ¿A partir de qué valor una familia debe pagar impuestos?**

Solución: Se tiene la variable I: «ingreso familiar anual», $X \sim \mathcal{N}(8000, 1200^2)$.

a) Debemos hallar la probabilidad $\Pr(I < 6600)$.

$$\begin{aligned}\Pr(I < 6600) &= F(6600) = \Phi\left(\frac{6600 - 8000}{1200}\right) \\ &= \Phi(-1.17) = 0.121.\end{aligned}$$

El 12.1 % de las familias tiene ingresos anuales menores a 6600 dólares. Eso quiere decir que son $0.121 \times 150\,000 = 18\,150$ familias.

b) Si buscamos el último quintil de ingreso, quiere decir aquellas familias que tienen el 20 % de los ingresos más altos; o sea, tenemos que encontrar un valor x de manera que $\Pr(I > x) = 0.2$.

$$\begin{aligned}\Pr(I > x) &= 1 - \Pr(I \leq x) = 0.2, \\ \Pr(I \leq x) &= F(x) = 1 - 0.2 = 0.8, \\ \Phi\left(\frac{x - 8000}{1200}\right) &= 0.8.\end{aligned}$$

En la tabla de la ley normal, vemos que se verifica que $\Phi(0.84) = 0.8$. Por tanto,

$$\frac{x - 8000}{1200} = 0.84.$$

Al resolver esta ecuación, nos da $x = 9008$. Consecuentemente, el 20 % de las familias tiene un ingreso superior a 9008 dólares anuales.

5. Se tomaron dos exámenes sobre 100 puntos, en el primero se obtuvo $\mu_1 = 80$, $\sigma_1 = 4$ y en el segundo $\mu_2 = 65$, $\sigma_2 = 5$. Un estudiante sacó 84 en el primer examen y 75 en el segundo. Comparativamente, ¿en cuál de los exámenes obtuvo mejor resultado?

Solución: Determinemos, para cada examen, el porcentaje de compañeros que sacaron menor nota que él, sabiendo que

$$\mu_1 = 80, \quad \sigma_1 = 4, \quad \mu_2 = 65, \quad \sigma_2 = 5.$$

$$\begin{aligned}\Pr(X < 84) &= F_1(84) = \Phi\left(\frac{84 - 80}{4}\right) = \Phi(1) = 0.8413, \\ \Pr(X < 75) &= F_2(75) = \Phi\left(\frac{75 - 65}{5}\right) = \Phi(2) = 0.9772.\end{aligned}$$

Como en el primer examen el porcentaje de compañeros que obtuvo menor nota es 84.13 % y en el segundo 97.72 %, tuvo, comparativamente mejor resultado en el segundo examen.

6. Una empresa embotella refrescos mediante una máquina que envasa el líquido, con un media μ y desviación estándar de $10\,\text{cm}^3$. Calcular el valor de la media para que solo se rebase la cantidad de $310\,\text{cm}^3$ en el 5 % de las botellas, si se supone que la cantidad de líquido embotellado tiene distribución normal.

Solución: Sea X: «la cantidad de líquido embotellado», con $X \sim \mathcal{N}(\mu, (10)^2)$. Se busca el valor de μ tal que

$$\Pr(X > 310) = 0.05.$$

Ahora bien,

$$\Pr(X > 310) = \Pr\left(\frac{X - \mu}{\sigma} > \frac{310 - \mu}{10}\right) = \Pr\left(Z > \frac{310 - \mu}{10}\right),$$

donde $Z \sim \mathcal{N}(0,1)$. Por la tabla de la ley normal, $\Pr(Z > 1.645) = 0.05$, debiéndose cumplir que

$$\frac{310-\mu}{10} = 1.645.$$

Así,

$$\mu = 310 - 10(1.645) = 293.55.$$

El marcador que indica la cantidad media de líquido debe estar posicionado en $293.5\,\text{cm}^3$. ◂

En el Cuadro 4.1 se encuentra un resumen de las leyes de probabilidad analizadas en este capítulo.

Distribución	Notación	E(X)	Var(X)
Uniforme discreta, n		$\frac{n+1}{2}$	$\frac{n^2-1}{12}$
Hipergeométrica	$\mathcal{H}(N,n,r)$	$\frac{rn}{N}$	$\frac{rn}{N}\left(1-\frac{n}{N}\right)\left(\frac{N-r}{N-1}\right)$
Bernoulli	Ber(p)	p	pq
Binomial	Bin(n,p)	np	npq
Geométrica	$\mathcal{G}(p)$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Binomial negativa	BN(r,p)	$\frac{r}{p}$	$r\frac{1-p}{p^2}$
Poisson	$\mathcal{P}(\lambda)$	λ	λ
Uniforme continua	$\mathcal{U}(a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponencial	$\mathcal{E}(\lambda)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Normal	$\mathcal{N}(\mu,\sigma^2)$	μ	σ^2
Normal estándar	$\mathcal{N}(0,1)$	0	1

Cuadro 4.1: Principales leyes de distribución de probabilidad

4.10. Ejercicios

Ley uniforme

1. Una variable aleatoria X tiene distribución uniforme sobre $[-3,1]$. Calcule:

   a) $\Pr(X=0)$;
   b) $\Pr(X<0)$;
   c) $\Pr(|X|<1)$;
   d) $\Pr(|X|>0.5)$
   e) Halle un valor de t tal que $\Pr(X>t) = \frac{1}{3}$.

2. Realice el ejercicio anterior considerando que $X \sim \mathcal{U}[-3,2]$.

3. Un reloj de manecillas se detuvo en un punto que no sabemos. Determine la probabilidad de que se haya detenido en los primeros 25 minutos luego de señalar la hora en punto.

4. Los autobuses de cierta línea salen con horario estricto cada cinco minutos. Halle la probabilidad de que un pasajero que llega a la parada tenga que esperar el autobús menos de tres minutos.

5. Al estudiar las ofertas de contratos de envío, un fabricante de computadoras ve que los contratos de los interesados tienen ofertas que se distribuyen uniformemente entre 20 mil y 25 mil dólares. Calcule la probabilidad de que el siguiente contrato sea:

   a) menor que 22 mil dólares;
   b) mayor que 24 mil dólares;
   c) Estime el costo medio de las ofertas en contratos de este tipo.

6. Supóngase que la velocidad de los autos en un sector de una carretera sigue una ley uniforme entre 60 y 120 km/h. ¿Cuál es la probabilidad de que un auto:

   a) tenga una velocidad de 80 km/h?;
   b) tenga una velocidad menor que 95 km/h?;
   c) tenga una velocidad menor que 70 km/h o mayor que 100 km/h?

7. El diámetro, x de un círculo se mide aproximadamente $5 \leq x \leq 6$ cm. Considerando el diámetro como una magnitud aleatoria X distribuida uniformemente en el intervalo (5, 6). Halle:

   a) la probabilidad de que el diámetro sea mayor que 5.8 cm;
   b) la esperanza matemática y la varianza del área del círculo.

8. Una llamada telefónica llegó a un conmutador en un tiempo, al azar, dentro de un periodo de un minuto. El conmutador estuvo ocupado durante 15 segundos en ese minuto. Calcule la probabilidad de que la llamada haya llegado mientras el conmutador no estuvo ocupado.

9. A partir de las 12:00 de la noche un centro de cómputo trabaja dos horas y para una. Una persona llama al centro un momento aleatorio entre las 12:00 y las 05:00 horas. ¿Cuál es la probabilidad de que el centro esté trabajando cuando llama la persona?

10. En una práctica de precisión aérea se deja caer una bomba a lo largo de una línea de un kilómetro de longitud. El blanco se encuentra en el punto medio de la línea. El blanco se destruirá si la bomba cae a una distancia menor que setenta y cinco metros del centro. Calcule la probabilidad de que el blanco se destruya si la bomba cae al azar a lo largo de la línea.

Ley exponencial

11. Escriba las funciones de densidad y distribución y los valores de la esperanza y la varianza para las variables aleatorias que siguen una ley exponencial:

    a) $\mathcal{E}(6)$; b) $\mathcal{E}(3)$; c) $\mathcal{E}(0.5)$; d) $\mathcal{E}(0.25)$.

12. Se prueban dos elementos que trabajan independientemente. El tiempo de trabajo del primer elemento tiene distribución $\mathcal{E}(0.02)$ y el segundo elemento $\mathcal{E}(0.05)$. Halle la probabilidad de que en el tiempo de duración $t = 6$ horas:

    a) ambos elementos fallen;
    b) ambos elementos no fallen;
    c) solo falle un elemento;
    d) falle por lo menos un elemento.

13. La duración (en minutos) de las llamadas telefónicas de larga distancia desde Quito es una variable aleatoria con densidad

$$f(t) = \begin{cases} 0, & \text{si } t \leq 0; \\ ce^{-t/3}, & \text{si } t > 0. \end{cases}$$

Determine el valor de c y calcule la probabilidad de que una llamada dure:

a) menos de 3 minutos;

b) más de 6 minutos;

c) entre 3 y 6 minutos.

d) Calcule la esperanza de la variable aleatoria e interprete su significado;

e) Si el costo del minuto de las llamadas telefónicas es de 40 centavos, ¿cuánto esperaría un usuario pagar por una llamada?

14. La duración (en años) de la vida de los individuos de una población humana se puede modelar mediante una variable aleatoria con función de densidad

$$f(t) = \begin{cases} \frac{1}{60}e^{-t/60}, & \text{si } t > 0; \\ 0, & \text{si } t \leq 0. \end{cases}$$

a) Determine la vida media de la población;

b) ¿Cuál es la probabilidad de que un individuo no llegue a los 42 años?;

c) ¿Cuál es la probabilidad de que una persona que tiene más de 50 años, supere los 65?

15. Suponga que la duración, en minutos, de una conversación telefónica sigue una ley exponencial $\mathcal{E}(1/5)$. Encuentre la probabilidad de que la duración de una conversación telefónica:

a) exceda los 5 min;

b) dure entre 3 y 6 min;

c) dure menos de 3 min;

d) dure menos de 6 min, dado que ha durado más de 3 min.

16. Se prueban tres elementos que trabajan independientemente entre sí. La duración del tiempo de trabajo sin fallo está distribuida según una ley exponencial: para el primer elemento $f_1(t) = 0.1e^{-0.1t}$, para el segundo elemento $f_2(t) = 0.2e^{-0.2t}$, para el tercer elemento $f_3(t) = 0.3e^{-0.3t}$. Halle la probabilidad de que en el intervalo de tiempo $(0, 10)$ horas, fallen:

a) por lo menos un elemento;

b) no menos de dos elementos.

17. La escala Richter para medir la magnitud de los terremotos sigue una ley exponencial de media 2.4. Calcule la probabilidad de que un sismo sea:

a) mayor que 3 grados en la escala de Richter;

b) entre 2 y 3 grados en la escala de Richter;

c) El sismo producido en la India el 30 de septiembre de 1993 tuvo la intensidad de 6.4 grados, ¿cuál es la probabilidad de que un sismo supere esta intensidad?

18. El tiempo de duración, en meses, de un tipo de resistencia eléctrica se expresa mediante una variable aleatoria X que sigue una ley exponencial $\mathcal{E}(0.5)$.

a) ¿Cuál es la probabilidad de que una de tales resistencias eléctricas dure más de 4 meses?

b) Si se prueban 10 resistencias eléctricas, ¿cuál es la probabilidad de que ninguna dure más de 4 meses?

c) ¿Cuántas resistencias se probarían para que con probabilidad igual a 0.9 se tenga al menos una resistencia que dure más de 4 meses?

d) Si el costo de producción de una resistencia es $C = 2 + (30 - X)^2$, ¿cuál es el costo esperado de una resistencia?

19. El tiempo T que se demora para completar una reparación eléctrica es una variable aleatoria distribuida exponencialmente, con media 10 horas. El costo C de llevar a cabo este trabajo se relaciona con el tiempo empleado mediante la fórmula

$$C = 100 + 40T + 3T^2.$$

a) Calcule el costo esperado de la reparación;

b) ¿Con qué frecuencia el tiempo será mayor que 20 horas?

20. La duración de los neumáticos de una marca determinada siguen una ley exponencial cuyo promedio es 30 (en miles de kilómetros). Calcule la probabilidad de que un neumático dure:

a) más de 30 mil km;

b) más de 30 mil km, dado que ha durado 15 mil km.

Ley normal

21. Se tiene una variable aleatoria Y con media 5 y varianza 16.

a) Determine su función de densidad.

b) Halle las probabilidades: $\Pr(Y < 6)$, $\Pr(Y > 4)$ y $\Pr(|Y| < 3)$.

22. Una variable aleatoria Z está distribuida normalmente, $Z \sim \mathcal{N}(1, 16)$. Calcule:

a) $\Pr(Z < 0)$; b) $\Pr(Z \geq 3)$; c) $\Pr(|Z| < 3)$; d) $\Pr(|Z| > 2)$.

23. Se sabe que el gasto en cigarrillos es, para los fumadores, de 5 dólares diarios por término medio, y que la desviación estándar es de 0.8 dólares. Suponiendo que el gasto sigue una distribución normal, ¿qué proporción de los fumadores gastan entre 4 y 6.2 dólares diarios?

24. Se experimenta con un medicamento que produce variación en el peso de las personas que lo toman. Pruebas de laboratorio han demostrado que al cabo de un mes la variación del peso sigue una distribución gaussiana de media 2 kg y desviación estándar 1.25 kg. Determine la probabilidad de que una persona:

a) haya aumentado al menos 1 kg;

b) haya rebajado de peso;

c) haya aumentado menos de 3 kg.

25. La compañía aérea Helios sabe que el tiempo de retraso de sus vuelos sigue una ley normal, con un retraso medio de 10 minutos y desviación estándar 5 minutos. Calcule la probabilidad de que:

a) un vuelo no tenga retraso;

b) el próximo vuelo llegue con no más de 12 minutos de retraso;

c) el próximo vuelo llegue con más de 15 minutos de retraso.

26. La Cruz Roja ha determinado que tiempo necesario para que una de sus ambulancias llegue al sitio donde hay una emergencia se distribuye según una variable normal de media 17 minutos y desviación estándar 3 minutos.

a) Calcule la probabilidad de que el tiempo de llegada esté comprendido entre 12 y 21 minutos;

b) ¿Para qué valor del tiempo t, la probabilidad de que la ambulancia emplee más de t minutos en llegar es del 5 %?

27. Los errores de la medición de peso de una balanza obedecen a una ley normal con desviación estándar 20 mg y esperanza 0 mg. Halle la probabilidad de que de tres mediciones independientes, el error de por lo menos una de ellas no sea mayor, en valor absoluto, que 4 mg.

28. Se aplicó una prueba de fluidez verbal a 500 alumnos de Educación Básica. Se supone que las puntuaciones obtenidas se distribuyen según una normal de media 80 y desviación estándar 12.

a) ¿Qué puntuación separa el 25 % de los alumnos con menos fluidez verbal?;

b) ¿A partir de qué puntuación se encuentra el 45 % de los alumnos con mayor fluidez verbal?;

c) ¿Cuántos alumnos tienen una fluidez menor que 76 puntos?

29. El perímetro craneal de los hombres, en medido en cm, es una variable aleatoria normal $\mathcal{N}(60, 4)$.

a) ¿Qué perímetro craneal debe tener un hombre para que el 16.6 % de sus paisanos «tengan más cabeza que él»?

b) ¿Y cuánto para que el 25.2 % tenga menos?

30. Se llama cociente intelectual (C.I.) al cociente entre la edad mental y la edad real. Se sabe que la ley de distribución del C.I. es normal con media 0.95 y desviación estándar 0.22. En una población con 2600 personas se desea saber:

a) ¿Cuántas tendrán un C.I. superior a 1.3?;

b) ¿Cuántas tendrán un C.I. inferior a 0.77?;

c) ¿Cuántas tendrán un C.I. entre 0.8 y 1.15?

31. Se va a construir un marco para montar una puerta. ¿Qué altura mínima ha de tener el marco para que el 1 % de la población tenga riesgo de chocar su cabeza al atravezarla, si la estatura de la población esta distribuida normalmente, con media $\mu = 1.72\,\text{m}$ y varianza σ^2, con $\sigma = 12\,\text{cm}$?

32. La estatura de la población masculina está normalmente distribuida con $\mu = 167\,\text{cm}$ y $\sigma = 3\,\text{cm}$.

a) ¿Cuál es la probabilidad de que un hombre tenga una estatura:

(i) mayor que 167 cm?; (ii) mayor que 170 cm?; (iii) entre 161 y 173 cm?

b) En una muestra aleatoria de cuatro hombres, ¿cuál es la probabilidad que:

(i) todos tengan estatura mayor que 170 cm?

(ii) dos tengan estatura menor que la media (y dos mayor que la media)?

33. El peso de las fundas de papas fritas producidas por una fábrica sigue una distribución normal con media 12.8 onzas y desviación estándar 0.6 onzas.

a) ¿Qué proporción de las fundas pesan más de 12 onzas?;

b) ¿Qué proporción de las fundas pesan entre 13 y 14 onzas?;

c) Determine el peso tal que el 12.5 % de las fundas pesen más que ese peso;

d) Si el fabricante desea mantener la media en 12.8 onzas, pero ajusta la desviación estándar tal que solo el 1 % de las fundas pese menos de 12 onzas, ¿cuál debe ser el valor de la desviación estándar?

34. La estatura de la población masculina y femenina siguen leyes de distribución normal. La masculina tiene $\mu_1 = 1.67\,\text{m}$ y $\sigma_1 = 12\,\text{cm}$ y la femenina $\mu_2 = 1.55\,\text{m}$ y $\sigma_2 = 10\,\text{cm}$. Se tiene una pareja en la cual el varón mide 1.70 m y la mujer 1.60 m. Comparativamente, ¿cuál de los dos es más alto respecto a los miembros de su sexo?

35. Los conductores que se fabrican para utilizar en las computadoras deben tener resistencias que varían entre 0.12 y 0.14 ohm. Las medidas de las resistencias que produce una compañía siguen una ley de distribución normal de media 0.13 ohm y desviación estándar 0.005 ohm.

    a) ¿Qué porcentaje de la producción de la compañía cumple con las especificaciones?;

    b) Si se usan cuatro de esos conductores en una computadora, ¿cuál es la probabilidad de que los cuatro cumplan con las especificaciones?

36. Los tiempos de la primera avería de una máquina de cierta marca tienen distribución gaussiana con un promedio de 1500 horas de uso y desviación estándar de 200 horas.

    a) ¿Qué fracción de esas máquinas fallarán antes de 1000 horas?;

    b) ¿Cuál debe ser el tiempo de garantía que deba dar el fabricante si desea que solo se presente el 5 % de las averías dentro del tiempo de garantía?

37. El promedio de las calificaciones de los estudiantes universitarios se distribuye normalmente con media 5.4 y desviación estándar igual a 0.5 puntos.

    a) ¿Qué porcentaje de los estudiantes tiene un promedio de calificaciones superior a 6?;

    b) Si los estudiantes que tienen un promedio inferior o igual a 4.9 abandonan la universidad, ¿qué porcentaje de alumnos desertará?;

    c) Se seleccionan al azar tres estudiantes, ¿cuál es la probabilidad de que los tres tengan un promedio de calificaciones superior a 6?

38. En el grupo étnico A, la estatura de las personas (en cm) sigue una distribución $\mathcal{N}(165; 25)$; en el grupo étnico B sigue una $\mathcal{N}(170; 25)$ y en el grupo C una $\mathcal{N}(175; 25)$. Los tres grupos étnicos son muy numerosos.

    a) Si elegimos una persona del grupo A, ¿cuál es la probabilidad de que mida más de 160 cm?;

    b) Si elegimos 10 personas al azar del grupo étnico A, independientemente unas de otras, ¿cuál es la probabilidad de que 5 de ellas midan más de 160 cm?;

    c) En una ciudad, el 50 % de la población pertenece a la etnia A, el 20 % pertenece a la B y el 30 % restante a la C. Si elegimos una persona al azar en esta ciudad y mide más de 172 cm, ¿cuál es la probabilidad de que pertenezca al grupo étnico C?;

    d) Si elegimos 10 personas al azar del grupo B, independientemente unas de otras, ¿cuál es la probabilidad de que al menos 5 midan más de 172 cm?

39. Una máquina para llenar cajas de cereal tiene una desviación estándar de 25 gramos sobre el peso de llenado de las cajas. ¿Qué medida debe indicar el marcador de llenado de las cajas para que permita que haya cajas de 450 gramos o más durante el 1 % del tiempo? Se supone que la cantidad de cereal por caja sigue una ley normal.

40. La anchura, en mm, de una población de coleópteros sigue una distribución $\mathcal{N}(\mu; \sigma^2)$. Se estima que el 77 % de la población mide menos de 12 mm y que el 84 % mide más de 7 mm. Halle los parámetros de la ley.

4.11. El teorema del Límite Central

Examinemos el siguiente teorema, que tiene una importancia fundamental, ya que constituye el nexo de comunicación entre las teorías de la probabilidad y la estadística.

Teorema (del Límite Central) Sean X_1, X_2, ..., X_n, n variables aleatorias independientes, distribuidas con media μ y varianza σ^2, y que siguen una ley de probabilidad cualquiera –no necesariamente la misma–. Se forma la variable suma

$$Y = X_1 + X_2 + \cdots + X_n,$$

que tiene esperanza $\mathbf{E}(Y) = n\mu$ y varianza $\mathrm{Var}(Y) = n\sigma^2$. Entonces, la distribución de la variable aleatoria

$$Z = \frac{Y - \mathbf{E}(Y)}{\sqrt{\mathrm{Var}(Y)}} = \frac{Y - n\mu}{\sigma\sqrt{n}}$$

tiende hacia una ley de distribución normal estándar, cuando n tiende al infinito.

El teorema implica que si n es grande, se puede aproximar las probabilidades de Y utilizando que

$$\Pr(Y \leq t) = \Pr\left(Z \leq \frac{t - n\mu}{\sigma\sqrt{n}}\right) \approx \Phi\left(\frac{t - n\mu}{\sigma\sqrt{n}}\right),$$

en la que Z es una variable aleatoria normal estándar.

En la práctica, se asume que la aproximación es buena si $n \geq 25$.

La formulación de este teorema es, en su forma más elemental, debida a P. S. Laplace y fue demostrado rigurosamente, en primer lugar, por Liapunov en 1901.

Ejemplos

1. Sean X_1, X_2, ..., X_{50}, cincuenta variables aleatorias independientes que siguen la ley

x	0	1	2
$\Pr(X = x)$	1/8	3/8	1/2

Calcule la probabilidad $\Pr(X_1 + X_2 + \cdots + X_{50} > 70)$.

Solución: La esperanza y la varianza de las variables aleatorias son

$$\mathbf{E}(X_i) = \frac{11}{8}, \quad \mathrm{Var}(X_i) = \frac{31}{64}.$$

Entonces, si $Y = X_1 + X_2 + \cdots + X_{50}$,

$$\begin{aligned} \mathbf{E}(Y) &= 50 \times \frac{11}{8} = \frac{275}{4}, \\ \mathrm{Var}(Y) &= 50 \times \frac{31}{64} = \frac{775}{32}. \end{aligned}$$

Por el Teorema del Límite Central,

$$
\begin{aligned}
\Pr(Y > 70) &= 1 - \Pr(Y \leq 70) = 1 - \Pr\left(Z \leq \frac{70 - \frac{275}{4}}{\sqrt{\frac{775}{32}}}\right) \\
&\approx 1 - \Phi\left(\frac{70 - \frac{275}{4}}{\sqrt{\frac{775}{32}}}\right) = 0.4013.
\end{aligned}
$$

2. El costo diario de operar un autobús tiene un costo fijo de 30 dólares y un valor variable del 30% de los ingresos. El ingreso tiene una distribución uniforme entre 50 y 250 dólares. a) Calcular la probabilidad de que el costo de operar un autobús, durante 31 días, supere los 2500 dólares; b) ¿Cuántos días de operación serán necesarios para que con una probabilidad de 0.95, el costo de operación sea de al menos 2350 dólares?

Solución: Definamos las siguientes variables aleatorias:

X: Ingreso diario por operación del autobús: $X \sim \mathcal{U}[50, 250]$.

C: Costo diario de operación del autobús: $C = 30 + 0.3X$.

Se tiene que

$$\mathbf{E}(X_i) = 150, \quad \mathbf{Var}(X_i) = 3333.33.$$

Entonces,

$$
\begin{aligned}
\mathbf{E}(C_i) &= 30 + 0.3 \times 150 = 75, \\
\mathbf{Var}(C_i) &= (0.3)^2 \times 3333.33 = 300.
\end{aligned}
$$

a) Sea $Y = \sum_{i=1}^{31} C_i$, el costo de operación mensual del autobús. Entonces,

$$
\begin{aligned}
\Pr(Y > 2500) &\approx 1 - \Phi\left(\frac{2500 - 31 \times 75}{\sqrt{300}\sqrt{31}}\right) = 1 - \Phi(1.815) \\
&= 1 - 0.9648 = 0.0352.
\end{aligned}
$$

b) Sea U el costo de operación en n días; entonces, la variable aleatoria

$$Z = \frac{U - 75n}{\sqrt{300}\sqrt{n}} \sim \mathcal{N}(0, 1).$$

Se debe determinar el valor de n, tal que $\Pr(U > 2350) = 0.95$:

$$\Pr(U > 2350) \approx \Pr\left(Z > \frac{2350 - 75n}{\sqrt{300}\sqrt{n}}\right) = 1 - \Phi\left(\frac{2350 - 75n}{\sqrt{300}\sqrt{n}}\right) = 0.95.$$

Por tanto,

$$\frac{2350 - 75n}{\sqrt{300}\sqrt{n}} = -1.645.$$

Al resolver esta ecuación, resulta que $n = 33.53$. Es decir, se necesitan 34 días.

3. Suponga que la vida útil de un componente electrónico se distribuye exponencialmente con media de 100 horas. Apenas falla un componente, se lo reemplaza con otro para continuar el trabajo. a) Calcular la probabilidad de que durante 210 días se necesiten más de 36 de esos componentes; b) ¿Cuántos de estos componentes se necesitan para que duren al menos 4600 horas, con una probabilidad del 99 %?

Solución: Definamos las siguientes variables aleatorias:

X_i: «Duración de los componentes electrónicos»; $X_i \sim \mathcal{E}(1/100)$ y $\mathbf{E}(X) = \sigma(X) = 100$.

Y_n: «Tiempo total de duración de n componentes»; $Y_n = \sum_{i=1}^{n} X_i$.

a) Se necesitan más de 36 componentes durante 210 días, si la vida útil es menor a 5040 horas (210 días por 24 horas). De manera que

$$\begin{aligned}\Pr(Y_{36} < 5040) &\approx \Phi\left(\frac{5040 - 36 \times 100}{100\sqrt{36}}\right) = \Phi(2.4) \\ &= 0.9918.\end{aligned}$$

b) Se debe encontrar n tal que $\Pr(Y_n \geq 4600) = 0.99$, luego

$$\Pr(Y_n < 4600) \approx \Phi\left(\frac{4600 - 100n}{100\sqrt{n}}\right) = 0.01.$$

Luego,

$$\frac{4600 - 100n}{100\sqrt{n}} = -2.33$$

La solución de la ecuación es $n = 64.5$; es decir, 65 componentes. ◀

Aproximación de la ley binomial por la normal

Un caso particular del Teorema del Límite Central –conocido como teorema de Moivre-Laplace–, se presenta cuando todas las variables X_i son independientes, idénticamente distribuidas según una ley de Bernoulli con parámetro p. Como sabemos, la variable

$$Y = \sum_{i=1}^{n} X_i$$

sigue una distribución $\mathbf{Bin}(n,p)$, con media np y varianza npq, con $q = 1 - p$. Por el Teorema del Límite Central, la variable

$$Z = \frac{Y - np}{\sqrt{npq}}$$

sigue aproximadamente una ley normal estándar, cuando n es suficientemente grande; es decir,

$$\Pr(Y \leq t) = \Pr\left(Z \leq \frac{t - np}{\sqrt{npq}}\right) \approx \Phi\left(\frac{t - np}{\sqrt{npq}}\right).$$

En la Figura 4.5 se muestra los valores de la distribución binomial para $n = 20$ y $p = 0.5$. Para esta distribución, $\mu = np = 20 \times 0.5 = 10$ y $\sigma^2 = np(1-p) = 20 \times 0.5 \times 0.5 = 5$. Sobrepuesta a la distribución binomial se encuentra una distribución normal con media $\mu = 10$ y varianza $\sigma^2 = 5$. Notamos que la curva de la normal se aproxima mucho al histograma de la binomial.

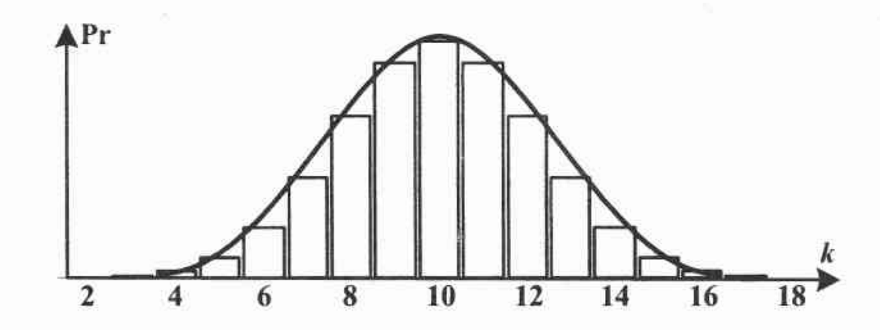


Figura 4.5: Aproximación de la ley binomial por la ley normal.

En la siguiente tabla se presenta una relación entre los parámetros n y p para que la aproximación normal a la ley binomial sea válida[4].

p	n requerido	p	n requerido
0+	221	0.25	74
0.01	214	0.30	51
0.05	188	0.35	32
0.10	157	0.40	16
0.15	128	0.45	13
0.20	100	0.50	13

Otro criterio para escoger n es que el intervalo $\left(p - 2\sqrt{\frac{pq}{n}},\, p + 2\sqrt{\frac{pq}{n}}\right)$ se encuentre completamente dentro del intervalo (0, 1).

Ejemplo. La Superitendencia de Bancos cree que el 32 % de los créditos al sector agrícola están en mora. En un estudio se tomo una muestra de 270 créditos a la agricultura. a) Hallar la probabilidad de que más de 80 de ellos estén en mora; b) ¿Cuál es la probabilidad de que exactamente 95 clientes estén en mora?

Solución: Sea X el número de clientes con créditos en mora; $X \sim \mathbf{Bin}(270, 0.32)$.

a) La probabilidad de que más de 80 clientes estén en mora es

$$\Pr(X > 80) = \Pr(X \geq 81) = \sum_{k=81}^{270} \mathbf{C}_{270}^{k}(0.32)^{k}(0.68)^{270-k},$$

cuyo cálculo puede ser muy complicado. Aplicando la aproximación de la ley normal a la ley binomial, se tiene

$$\mu = np = 270 \times 0.32 = 86.4, \quad \sigma = \sqrt{npq} = \sqrt{270 \times 0.32 \times 0.68} = 7.665.$$

Luego,

$$\begin{aligned}\Pr(X > 80) &= \Pr\left(Z > \frac{80 - 86.4}{7.665}\right) \approx 1 - \Phi\left(\frac{80 - 86.4}{7.665}\right) \\ &= 1 - 0.2033 = 0.7967.\end{aligned}$$

[4] Una amplia discusión del tema se encuentra en Samuels, M. y Lu, T.-F.C. (1992), "Sample Size Requirements for the Back-of-the-Envelope Binomial Confidence Interval," *The American Statistician*, 46, 228-231, de la cual se extrajo la tabla.

b) La probabilidad buscada es $\Pr(X = 95) = C_{270}^{95}(0.32)^{95}(0.68)^{270-95} = 0.02735$.

Si empleamos el teorema de Moivre-Laplace, resulta:

$$\begin{aligned}\Pr(X = 95) &= \Pr(94.5 \leq X \leq 95.5) \approx \Phi\left(\frac{95.5 - 86.4}{7.665}\right) - \Phi\left(\frac{94.5 - 86.4}{7.665}\right)\\ &= 0.8829 - 0.8554 = 0.0275.\end{aligned}$$

Como se observa, la diferencia entre el valor exacto y el aproximado es mínima.

4.12. Ejercicios

1. En una caja se empacan 100 latas de conservas. Según los datos de la fábrica, cada lata tiene un peso promedio de 1 oz con desviación estándar de 0.1 oz. ¿Cuál es la probabilidad de que una caja pese más de 102 oz?

2. Un borracho camina de forma aleatoria de la siguiente forma: cada minuto da un paso hacia adelante o hacia atrás con igual probabilidad y con independencia de los pasos anteriores. Cada paso es de 50 cm. Calcule la probabilidad de que en una hora avance más de 5 metros.

3. Los clientes de cierto banco efectúan depósitos con media 157.92 dólares y desviación estándar 30.20 dólares. Aparte de ésto no se sabe nada más acerca de la distribución de estos depósitos. Como parte de un estudio, se eligieron al azar e independientemente 75 depósitos. ¿Cuál es la probabilidad de que la suma de estos 75 depósitos sea 12 750 dólares o mayor?

4. Los vehículos que cruzan un puente tienen pesos cuya media es de 4675 kg y cuya desviación estándar es de 345 kg. Si hay 40 vehículos sobre el puente en un instante dado, hallar el número a tal que la probabilidad (aproximada) de que su peso total no supere a a sea del 99 %.

5. La empresa Rapid Express envía paquetes de distintos pesos, con una media de 1.5 kg y una desviación estándar de 1.0 kg. Teniendo en cuenta que los paquetes provienen de una gran cantidad de clientes diferentes, es razonable modelizar sus pesos como variables aleatorias independientes. Calcule la probabilidad de que el peso total de 100 paquetes exceda de 170 kg.

6. El propietario de una copiadora ha determinado que el número diario de copias que se realizan en su local tiene una media de 1250 con una desviación estándar de 350. Halle la probabilidad de que en un mes de trabajo (25 días) el total de copias:

   a) sea menor a 30 000;
   b) se encuentre entre 25 000 y 32 000.

7. Una radióloga que trabaja en el servicio de traumatología de un hospital ha comprobado que el tiempo, en minutos, que tarda en atender a cada paciente es una variable aleatoria con media 7 y desviación estándar 2. Durante su jornada laboral trabaja 6 horas atendiendo pacientes sucesivamente y sin interrupción. Calcule, aproximadamente, la probabilidad de que durante un día pueda atender hasta 55 pacientes dentro del horario de su jornada laboral. Se supone que todos los pacientes están en la consulta con suficiente antelación y que no hay «tiempos vacíos» entre dos pacientes consecutivos.

8. La resistencia de un hilo metálico es una variable aleatoria cuya media es 3 kg y su desviación estándar 1 kg. Suponiendo que la resistencia de un cable es igual a la suma de las resistencias de los hilos que lo forman.

   a) Calcule la probabilidad de que un cable de 100 hilos sostenga 280 kg;

b) ¿Cuántos hilos se necesitan para que el cable sostenga 300 kg con un 99 % de seguridad?

9. Un jugador de baloncesto encesta un lanzamiento de 3 puntos con probabilidad 0.3.

   a) Aproxime la distribución del número de canastas conseguidas en 25 lanzamientos;
   b) Calcule la probabilidad de encestar más de 10 canastas.

10. En promedio, de las personas que ingresan a una librería solo el 25 % realiza una compra. Si en un día entraron 80 clientes, calcule la probabilidad aproximada de que se hagan al menos 28 compras.

11. Se ha encontrado que el 70 % de las personas que entran en un centro comercial realizan cuando menos una compra. Para una muestra de 50 personas,

    a) ¿cuál es la probabilidad de que cuando menos 40 de ellas realicen una ó más compras?;
    b) ¿cuál es la probabilidad de que menos de 30 de entre 50 personas muestreadas realicen cuando menos una compra?

12. En una fabrica microcircuitos se ha comprobado que el 4 % de estos son defectuosos. Un cliente compra un paquete de 500 microcircuitos procedentes de la fábrica. Determine:

    a) el número esperado de microcircuitos no defectuosos;
    b) la probabilidad de que se encuentre más de 25 microcircuitos defectuosos;
    c) la probabilidad de que el número de microcircuitos defectuosos esté entre 16 y 30.

13. Se conoce, por estudios previos, que la proporción de vacas que enfermarán después de suministrarles la vacuna contra la fiebre aftosa es del 2 %. Una granja tiene 600 vacas que son vacunadas. Determine:

    a) el número esperado de animales que no enfermarán;
    b) la probabilidad de que el número de reses que enferman sea, como máximo, 17;
    c) la probabilidad de que el número de reses que no enferman sea, como mínimo, 590.

14. Un zoólogo estudia cierta especie de ratones de campo. Para ello captura ejemplares de una población grande en la que el porcentaje de dicha especie es $100p$.

    a) Si $p = 0.3$, halle la probabilidad de que en 6 ejemplares capturados haya al menos 2 de los que le interesan;
    b) Si $p = 0.05$, calcule la probabilidad de que en 200 haya exactamente 6 de los que le interesan;
    c) Si $p = 0.4$, calcule la probabilidad de que en 200 haya entre 75 y 110 de los que le interesan.

15. Sea S_{45} el número de fracasos que preceden al 45° éxito en un proceso de Bernoulli con probabilidad de éxito $p = 0.36$. Sea $S_{45} = X_1 + \cdots + X_{45}$, donde X_1 representa el número de fracasos que preceden al primer éxito, X_2 es el número de fracasos entre el primer y el segundo éxito, y así sucesivamente. Las X_j son independientes.

    a) Dé el nombre de la distribución de una única X_j, obtenga también su media y su varianza;
    b) ¿Cuál es el valor esperado y la varianza de S_{45}?;
    c) Aproxime la probabilidad de que S_{45} esté a una distancia no superior a 20 de su media.

Capítulo 5

Distribuciones Multidimensionales de Probabilidad

El supuesto erróneo de que la correlación implica causalidad es probablemente uno de los dos o tres errores más serios y comunes del pensamiento humano
Stephen Jay Gould

En muchos casos, un fenómeno aleatorio no depende de una sola variable, sino de dos o más; por ejemplo, algo tan simple como el tiempo que empleamos en trasladarnos desde la casa a la universidad depende, entre otras cosas, de la velocidad media del carro y del número de veces que nos detengamos por los semáforos en luz roja. Es decir, el resultado en la prueba descrita depende de, al menos, dos variables aleatorias.

En lo que sigue, trataremos con conjuntos de varias variables aleatorias que se manifiestan simultáneamente en un fenómeno y determinaremos si ellas están o no relacionadas. Para simplificar la exposición, analizaremos el caso bivariante ya que la generalización a más variables es inmediata.

5.1. Variables aleatorias bidimensionales

Definición (de variable aleatoria bidimensional) Sean X y Y dos variables aleatorias unidimensionales definidas sobre un mismo espacio muestral Ω; entonces, la función

$$\begin{aligned} Z:\Omega &\longrightarrow \mathbf{R}\times\mathbf{R} \\ \omega &\longmapsto (X(\omega),Y(\omega)), \end{aligned}$$

donde ω es un evento elemental, es una variable aleatoria bidimensional.

Notación. Para referirnos a los $\omega \in \Omega$ tales que $X(\omega) = a$ y $Y(\omega) = b$, simplemente lo haremos como $X = a$, $Y = b$ y a la probabilidad de este suceso como $\Pr(X = a, Y = b)$.

De forma análoga se definen y notan las probabilidades $\Pr(X \leq a, Y \leq b)$ y $\Pr(a < X \leq c, b < Y \leq d)$.

5.1.1. Variables bidimensionales discretas

Definición (de variable aleatoria bidimensional discreta) Sean X y Y dos variables aleatorias discretas. La función de probabilidad conjunta de X y Y está dada por

$$f(x,y) = \Pr(X = x, Y = y).$$

A la variable aleatoria (X,Y) se le denomina bidimensional discreta.

Sea $T = \{(x,y) \in \mathbf{R}^2 / f(x,y) > 0\}$; es decir, el conjunto de puntos con probabilidad positiva, es finito o infinito numerable y se cumple que

$$\sum_{(x,y)\in T} f(x,y) = 1.$$

Supongamos que x_1, x_2, ... y y_1, y_2, ... son los valores posibles de X y Y, respectivamente, y sea

$$p_{ij} = \Pr(X = x_i, Y = y_j).$$

La probabilidad del evento $(X,Y) \in E$ es igual a la suma de todos los p_{ij} para los cuales $(x_i, y_j) \in E$:

$$\Pr[(X,Y) \in E] = \sum_{(x_i,y_j)\in E} \Pr(X = x_i, Y = y_j) = \sum_{(x_i,y_j)\in E} p_{ij}.$$

Definición (de función de distribución conjunta) La función de distribución conjunta de la variable aleatoria bidimensional (X,Y) se define por

$$F(x,y) = \Pr(X \le x, Y \le y) = \sum_{x_i \le x} \sum_{y_j \le y} p_{ij}.$$

La función de distribución conjunta cumple que $\lim_{x\to-\infty} \lim_{y\to-\infty} F(x,y) = 0$ y que $\lim_{x\to\infty} \lim_{y\to\infty} F(x,y) = 1$.

A partir de las p_{ij} se pueden encontrar las funciones de distribución de X y de Y.

Definición (de función de probabilidad marginal) Las funciones de probabilidad marginal de X y de Y están dadas por las fórmulas

$$\begin{aligned} f_X(x_i) &= \sum_{j=1}^{\infty} \Pr(X = x_i, Y = y_j) = \sum_{j=1}^{\infty} p_{ij}. \\ f_Y(y_j) &= \sum_{i=1}^{\infty} \Pr(X = x_i, Y = y_j) = \sum_{i=1}^{\infty} p_{ij}. \end{aligned}$$

Las funciones de distribución marginal F_X y F_Y se calculan por

$$F_X(t) = \sum_{x_i \le t} \sum_{j=1}^{\infty} p_{ij} \quad \text{y} \quad F_Y(t) = \sum_{y_j \le t} \sum_{i=1}^{\infty} p_{ij}.$$

Observación. Si los espacios muestrales son finitos, las series deben reemplazarse por sumas finitas.

Con ayuda de variables aleatorias bidimensionales se puede dar una definición de independencia equivalente a la anteriormente dada:

Definición (de independencia) Las variables aleatorias discretas X y Y son independientes si

$$\Pr(X = x, Y = y) = \Pr(X = x) \times \Pr(Y = y).$$

Ejemplos

1. Determine el valor de k para que la función

$$f(x,y) = kxy, \quad \text{para} \quad x = 1,2,3;\ y = 1,2,3,$$

pueda servir como una función conjunta de probabilidad.

Solución: Sustituyendo los valores de x y de y, encontramos que:

$$\begin{array}{lll} f(1,1) = k, & f(1,2) = 2k, & f(1,3) = 3k, \\ f(2,1) = 2k, & f(2,2) = 4k, & f(2,3) = 6k, \\ f(3,1) = 3k, & f(3,2) = 6k, & f(3,3) = 9k. \end{array}$$

Para que f sea una función de probabilidad, la suma de todos los términos que acabamos de calcular deben dar 1; es decir,

$$k + 2k + 3k + 2k + 4k + 6k + 3k + 6k + 9k = 1.$$

Resolviendo esta ecuación, resulta que $k = \frac{1}{36}$.

Pongamos en una tabla la función de probabilidad resultante:

	X		
Y	1	2	3
1	1/36	2/36	3/36
2	2/36	4/36	6/36
3	3/36	6/36	9/36

2. Las probabilidades de la distribución conjunta de las variables aleatorias S y T se dan por

	S		
T	-1	0	1
-1	1/8	1/12	7/24
1	5/24	1/6	1/8

La intersección de las filas y las columnas da la probabilidad $p_{ij} = \Pr(S = i, T = j)$ $(i = -1, 0, 1;\ j = -1, 1)$. a) Calcular $\Pr(S \leq 0.5, T \leq 0.3)$; b) Hallar las leyes de las variables aleatorias S y T.

Solución:

a) De acuerdo a la fórmula de la función de distribución conjunta de S y T:

$$F(s,t) = \Pr(S \leq s, T \leq t) = \sum_{i \leq s} \sum_{j \leq t} p_{ij}.$$

Entonces,

$$\begin{aligned} F(0.5, 0.3) &= \sum_{i \leq 0.5} \sum_{j \leq 0.3} \Pr(S = i, T = j) \\ &= \Pr(S = -1, T = -1) + \Pr(S = 0, T = -1) \\ &= \frac{1}{8} + \frac{1}{12} = \frac{5}{24}. \end{aligned}$$

b) Por la fórmula de la función de probabilidad marginal tenemos: $f_S(i) = \sum_j \Pr(S = i, T = j)$; por lo que

$$\begin{aligned} f_S(-1) &= \Pr(S = -1, T = -1) + \Pr(S = -1, T = 1) \\ &= \frac{1}{8} + \frac{5}{24} = \frac{1}{3}. \\ f_S(0) &= \Pr(S = 0, T = -1) + \Pr(S = 0, T = 1) \\ &= \frac{1}{12} + \frac{1}{6} = \frac{1}{4}. \\ f_S(1) &= \Pr(S = 1, T = -1) + \Pr(S = 1, T = 1) \\ &= \frac{7}{24} + \frac{1}{8} = \frac{5}{12}. \end{aligned}$$

De manera análoga, se obtiene la ley de T:

$$\begin{aligned} f_T(-1) &= \Pr(S = -1, T = -1) + \Pr(S = 0, T = -1) + \Pr(S = 1, T = -1) \\ &= \frac{1}{8} + \frac{1}{12} + \frac{7}{24} = \frac{1}{2}. \\ f_T(1) &= \Pr(S = -1, T = 1) + \Pr(S = 0, T = 1) + \Pr(S = 1, T = 1) \\ &= \frac{5}{24} + \frac{1}{6} + \frac{1}{8} = \frac{1}{2}. \end{aligned}$$

Entonces, las variables aleatorias S y T siguen las leyes:

S	-1	0	1
$\Pr(S = i)$	$1/3$	$1/4$	$5/12$

T	-1	1
$\Pr(T = j)$	$1/2$	$1/2$

5.1.2. Variables bidimensionales continuas

Definición (de variable aleatoria bidimensional continua) Sean X y Y dos variables aleatorias continuas, a la variable aleatoria (X, Y) se le denomina bidimensional continua si para cualquier punto $(x, y) \in \mathbf{R}^2$ se cumple

$$\Pr(X = x, Y = y) = 0.$$

A la variable aleatoria (X, Y) está asociada una función no negativa f, denominada *función de densidad conjunta*, que cumple con las siguientes propiedades:

1. $f(x, y)$ es no negativa; es decir, $\forall (x, y) \in \mathbf{R}^2$, $f(x, y) \geq 0$.
2. $\displaystyle\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dy\, dx = 1.$
3. La probabilidad del evento «$(X, Y) \in E$» se calcula mediante

$$\Pr[(X, Y) \in E] = \iint_E f(x, y)\, dy\, dx.$$

Definición (de función de distribución conjunta) La función de distribución conjunta de la variable aleatoria bidimensional (X, Y) se define por

$$F(x,y) = \Pr(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(s,t)\, dt\, ds.$$

La función de distribución conjunta cumple que $\lim\limits_{x\to-\infty} \lim\limits_{y\to-\infty} F(x,y) = 0$ y que $\lim\limits_{x\to\infty} \lim\limits_{y\to\infty} F(x,y) = 1$.

La función de densidad conjunta puede obtenerse a partir de la función de distribución mediante

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x\, \partial y}.$$

Si $E = [a,b] \times [c,d]$, la probabilidad del evento $(X,Y) \in E$ es igual a

$$\Pr(a \leq X \leq b, c \leq Y \leq d) = \iint_E f(x,y)\, dx\, dy = \int_c^d \int_a^b f(x,y)\, dx\, dy.$$

Definición (de función de densidad marginal) Las funciones de densidad marginal de las variables aleatorias X y Y están dadas, respectivamente, por las relaciones

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\, dy \quad \text{y} \quad f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\, dx.$$

A partir de éstas, se calculan las funciones de distribución marginal F_X y F_Y:

$$F_X(t) = \int_{-\infty}^{t} \int_{-\infty}^{\infty} f(x,y)\, dy\, dx \quad \text{y} \quad F_Y(t) = \int_{-\infty}^{t} \int_{-\infty}^{\infty} f(x,y)\, dx\, dy$$

Con este tipo de variables aleatorias también se puede reformular la definición de independencia.

Definición (de independencia) Las variables aleatorias continuas X y Y son independientes si para todo par de valores $(x,y) \in \mathbf{R}^2$ se cumple

$$F(x,y) = F_X(x)\, F_Y(y),$$

o equivalentemente,

$$f(x,y) = f_X(x)\, f_Y(y).$$

Ejemplos

1. Un círculo de radio a está inscrito dentro de un cuadrado cuyo lado tiene una longitud de $2a$ (véase la Figura 5.1). Se supone que la probabilidad de que un dardo arrojado hacia el cuadrado es idéntica para cualquier punto. a) Calcular la probabilidad de que el dardo impacte dentro del círculo; b) Encontrar las leyes marginales de X y de Y.

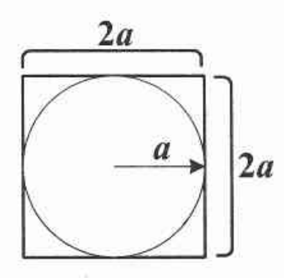


Figura 5.1:

Solución:

a) La probabilidad de impacto esta definida por la densidad

$$f_{XY}(x,y) = \begin{cases} \dfrac{1}{4a^2}, & \text{si } \ (x,y) \in [-a,a] \times [-a,a]; \\ 0, & \text{si } \ (x,y) \notin [-a,a] \times [-a,a]. \end{cases}$$

La probabilidad de impacto dentro del círculo $x^2 + y^2 = a^2$ se calcula como

$$\Pr(X^2 + Y^2 \leq a^2) = \iint_{x^2+y^2=a^2} f(x,y)\, dx\, dy = \frac{\pi a^2}{4a^2} = \frac{\pi}{4}.$$

b) Por simetría, las funciones de densidad de X y de Y son idénticas; calculemos la de X:

$$f_X(x) = \int_{-a}^{a} \frac{1}{4a^2}\, dy = \frac{1}{2a}, \quad x \in [-a,a].$$

2. Dada la función de densidad conjunta de la variable aleatoria bidimensional (X,Y)

$$f_{XY}(x,y) = \begin{cases} \dfrac{1}{x^2y^2}, & \text{si } x \geq 1,\ y \geq 1; \\ 0, & \text{caso contrario.} \end{cases}$$

Determinar: a) las funciones de densidad marginal de cada una de las variables; b) la función de distribución asociada.

Solución:

a) Las funciones de densidad marginal son

$$\begin{aligned} f_X(x) &= \int_1^{\infty} f_{XY}(x,y)\, dy = \int_1^{\infty} \frac{1}{x^2y^2}\, dy = \frac{1}{x^2}. \\ f_Y(y) &= \int_1^{\infty} f_{XY}(x,y)\, dx = \int_1^{\infty} \frac{1}{x^2y^2}\, dx = \frac{1}{y^2}. \end{aligned}$$

Adicionalmente, podemos deducir que las variables aleatorias son independientes.

b) La función de distribución es

$$\begin{aligned} F(x,y) &= \int_1^{y}\int_1^{x} f_{XY}(s,t)\, ds\, dt = \int_1^{y}\int_1^{x} \frac{1}{s^2t^2}\, ds\, dt \\ &= \left(\frac{1-x}{x}\right)\left(\frac{1-y}{y}\right). \end{aligned}$$

Consecuentemente, la función de distribución queda como

$$F_{XY}(x,y) = \begin{cases} \dfrac{(1-x)(1-y)}{xy}, & \text{si } x \geq 1, \quad y \geq 1; \\ 0, & \text{caso contrario.} \end{cases}$$

5.2. Distribución condicionada

Definición (de probabilidad condicionada) Sean a y b dos números reales cualesquiera y X y Y dos variables aleatorias de manera que $\Pr(Y \leq b) \neq 0$. La probabilidad condicionada de que $X \leq a$, dado que $Y \leq b$, se representa mediante $\Pr(X \leq a|Y \leq b)$, y se define mediante la igualdad

$$\Pr(X \leq a|Y \leq b) = \frac{\Pr(X \leq a, Y \leq b)}{\Pr(Y \leq b)}$$

- Para variables aleatorias discretas, la probabilidad condicionada de x, para un valor fijo de la variable y, está dada por
$$\Pr(X = x|Y = y) = \frac{\Pr(X = x, Y = y)}{\Pr(Y = y)}.$$
- Para variables aleatorias continuas, la función de densidad condicionada de x, para un valor fijo de la variable y, se calcula por
$$f(x|y) = \frac{f(x,y)}{f_Y(y)}.$$
Como $f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx = \int_{-\infty}^{\infty} f(y|x) f_X(x)\,dx$; entonces,
$$f(x|y) = \frac{f(y|x)f_X(x)}{\displaystyle\int_{-\infty}^{\infty} f(y|x)f_X(x)\,dx},$$
que puede interpretarse como el teorema de Bayes para funciones de densidad.

Ejemplos

1. **(Continuación)** Un vector aleatorio bidimensional sigue la ley:

	X		
Y	1	2	3
1	1/36	2/36	3/36
2	2/36	4/36	6/36
3	3/36	6/36	9/36

Hallar la distribución condicionada de X cuando $Y = 2$.

Solución: La distribución marginal de Y es

Y	1	2	3
$\Pr(Y = i)$	1/6	1/3	1/2

La distribución condicionada de X cuando $Y = 2$ será:

$$\begin{aligned}
\Pr(X = 1|Y = 2) &= \frac{\Pr(X = 1, Y = 2)}{\Pr(Y = 2)} = \frac{\frac{2}{36}}{\frac{1}{3}} = \frac{1}{6},\\
\Pr(X = 2|Y = 2) &= \frac{\Pr(X = 2, Y = 2)}{\Pr(Y = 2)} = \frac{\frac{4}{36}}{\frac{1}{3}} = \frac{1}{3},\\
\Pr(X = 3|Y = 2) &= \frac{\Pr(X = 3, Y = 2)}{\Pr(Y = 2)} = \frac{\frac{6}{36}}{\frac{1}{3}} = \frac{1}{2}.
\end{aligned}$$

2. La ley de densidad conjunta de una variable aleatoria bidimensional (X,Y) es

$$f_{XY}(x,y) = \begin{cases} 2, & \text{si } 0 \le x \le 1,\ 0 \le y \le 1,\ x+y \le 1; \\ 0, & \text{caso contrario.} \end{cases}$$

Hallar la ley de distribución condicional de Y cuando $X = x_0$.

Solución: Se tiene que

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx = \int_0^{1-y} 2\,dx = 2(1-y).$$

La distribución condicional de X cuando $Y = y_0$ es

$$f(x|y_0) = \frac{f(x,y_0)}{f_Y(y_0)} = \begin{cases} \dfrac{2}{2(1-y_0)} = \dfrac{1}{1-y_0}, & \text{para } 0 \le x \le 1 - y_0; \\ 0, & \text{caso contrario.} \end{cases}$$

5.3. Esperanza y covarianza de una variable aleatoria bidimensional

Al igual que en el caso de las variables aleatorias unidimensionales, en las bidimensionales es posible calcular la esperanza y la varianza, previa la realización de una transformación de variables.

Definición (de esperanza) Sean (X,Y) un vector aleatorio bidimensional y $g(x,y)$ una función real

$$\begin{aligned} g : \mathbf{R}^2 &\longrightarrow \mathbf{R} \\ (x,y) &\longmapsto g(x,y). \end{aligned}$$

1. Si (X,Y) es un vector aleatorio discreto, cuya función de probabilidad es $f(x,y)$, entonces

$$\mathbf{E}(g(X,Y)) = \sum_{x_i}\sum_{y_j} g(x_i,y_j)\, f(x_i,y_j) = \sum_{x_i}\sum_{y_j} g(x_i,y_j)\, p_{ij}.$$

2. Si (X,Y) es un vector aleatorio continuo, cuya función de densidad conjunta es $f(x,y)$, entonces

$$\mathbf{E}(g(X,Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\, f(x,y)\,dy\,dx.$$

Observemos que si X y Y son independientes, se deduce que

$$\mathbf{E}[g(X)h(Y)] = \mathbf{E}[g(X)]\mathbf{E}[h(Y)].$$

Para las variables aleatorias bidimensionales se tiene una medida estadística nueva, la covarianza, que permite evaluar la relación entre las variables aleatorias X y Y.

Definición (de covarianza) Sean X y Y dos variables aleatorias, la covarianza entre X y Y se calcula por

$$\mathrm{Cov}(X,Y) = \mathbf{E}[(X - \mathbf{E}(X))(Y - \mathbf{E}(Y))].$$

Equivalentemente, la covarianza se puede calcular como $\mathrm{Cov}(X,Y) = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y)$.

Propiedades. (Solo se demostrará una de ellas, se recomienda al lector verificar las restantes)

1. $\text{Cov}(X,Y) = \text{Cov}(Y,X)$.

2. $\text{Cov}(X,X) = \text{Var}(X)$.

3. Si a_1 y a_2 son dos constantes positivas, entonces

$$\text{Cov}(a_1X_1 + a_2X_2, Y) = a_1\,\text{Cov}(X_1,Y) + a_2\,\text{Cov}(X_2,Y).$$

4. Si las variables aleatorias son independientes, la covarianza entre ellas es igual a cero.

   En efecto,

$$\text{Cov}(X,Y) = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y) = \mathbf{E}(X)\mathbf{E}(Y) - \mathbf{E}(X)\mathbf{E}(Y) = 0.$$

5. $|\,\text{Cov}(X,Y)| \leq \sqrt{\text{Var}(X)\,\text{Var}(Y)}$.

A continuación, se deduce una expresión para la varianza de la suma de dos variables aleatorias cualesquiera.

$$\begin{aligned}
\text{Var}(X+Y) &= \mathbf{E}(X+Y)^2 - [\mathbf{E}(X+Y)]^2 \\
&= \mathbf{E}(X^2) + \mathbf{E}(Y^2) + 2\mathbf{E}(XY) - [\mathbf{E}^2(X) + \mathbf{E}^2(Y) + 2\mathbf{E}(X)\mathbf{E}(Y)] \\
&= [\mathbf{E}(X^2) - \mathbf{E}^2(X)] + [\mathbf{E}(Y^2) - \mathbf{E}^2(Y)] + 2[\mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y)] \\
&= \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X,Y).
\end{aligned}$$

Así, para dos variables aleatorias cualesquiera se tiene que

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X,Y).$$

De manera similar, la varianza del producto de dos variables aleatorias X y Y es

$$\begin{aligned}
\text{Var}(XY) &= \mu_Y^2\,\text{Var}(X) + \mu_X^2(X)\,\text{Var}(Y) + 2\mu_X\mu_Y\,\text{Cov}(X,Y) + 2\mu_X\mathbf{E}\left[(X-\mu_X)(Y-\mu_Y)^2\right] \\
&\quad +2\mu_Y\mathbf{E}\left[(Y-\mu_Y)(X-\mu_X)^2\right] + 2\mathbf{E}\left[(X-\mu_X)^2(Y-\mu_Y)^2\right] - [\text{Cov}(X,Y)]^2,
\end{aligned}$$

donde $\mu_X = \mathbf{E}(X)$ y $\mu_Y = \mathbf{E}(Y)$.

Con base en la covarianza y la varianza se define el coeficiente de correlación, que es una medida de la dependencia entre las variables aleatorias X y Y.

Definición (de coeficiente de correlación) Sean X y Y dos variables aleatorias, el coeficiente de correlación entre X y Y se calcula por

$$\rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}.$$

Propiedades. (Se recomienda que el lector verifique alguna de ellas)

1. $\rho(X,Y) = \rho(Y,X)$.

2. El valor del coeficiente de correlación varía entre -1 y 1; es decir, $-1 \leq \rho(X,Y) \leq 1$.

3. Si Y se expresa linealmente en función de X, por $Y = aX + b$, donde a y b son dos constantes, entonces $|\rho(X,Y)| = 1$.

4. Si $|\rho(X,Y)| = 1$, entonces existe dependencia lineal entre X y Y.

5. Si X y Y son variables aleatorias independientes, entonces $\rho(X,Y) = 0$.

Observación. Se debe tener en cuenta que si dos variables aleatorias son independientes, entonces son no correlacionadas; pero la afirmación recíproca no es correcta; es decir, si dos variables aleatorias no están correlacionadas, no son obligatoriamente independientes.

Ejemplos

1. Las variables aleatorias S y T tienen función de probabilidad conjunta dada por

	S		
T	-1	0	1
-1	$1/8$	$1/12$	$7/24$
1	$5/24$	$1/6$	$1/8$

Calcular: a) las esperanzas y las varianzas de S y de T; b) el coeficiente de correlación.

Solución:

a) Anteriormente habíamos calculado las leyes marginales de S y de T, que son:

S	-1	0	1
$\Pr(S=i)$	$1/3$	$1/4$	$5/12$

T	-1	1
$\Pr(T=j)$	$1/2$	$1/2$

Sus correspondientes esperanzas son

$$\begin{aligned}
\mathbf{E}(S) &= (-1)\left(\frac{1}{3}\right) + (0)\left(\frac{1}{4}\right) + (1)\left(\frac{5}{12}\right) = \frac{1}{12},\\
\mathbf{E}(S^2) &= (-1)^2\left(\frac{1}{3}\right) + (0)^2\left(\frac{1}{4}\right) + (1)^2\left(\frac{5}{12}\right) = \frac{9}{12},\\
\mathbf{E}(T) &= (-1)\left(\frac{1}{2}\right) + (1)\left(\frac{1}{2}\right) = 0,\\
\mathbf{E}(T^2) &= (-1)^2\left(\frac{1}{2}\right) + (1)^2\left(\frac{1}{2}\right) = 1.
\end{aligned}$$

En consecuencia,

$$\begin{aligned}
\mathrm{Var}(S) &= \mathbf{E}\left(S^2\right) - \mathbf{E}^2(S) = \frac{9}{12} - \left(-\frac{1}{12}\right)^2 = \frac{107}{144},\\
\mathrm{Var}(T) &= \mathbf{E}\left(T^2\right) - \mathbf{E}^2(T) = 1 - 0^2 = 1.
\end{aligned}$$

b) Calculemos, en primer lugar, la covarianza entre S y T, para ello determinemos $\mathbf{E}(ST)$:

$$\begin{aligned}
\mathbf{E}(ST) &= \sum_i \sum_j i\,j\,p(i,j) \quad (i = -1, 0, 1;\ j = -1, 1)\\
&= (-1)(-1)\left(\frac{1}{8}\right) + (-1)(1)\left(\frac{5}{24}\right) + (0)(-1)\left(\frac{1}{12}\right)\\
&\quad +(0)(1)\left(\frac{1}{6}\right) + (1)(-1)\left(\frac{7}{24}\right) + (1)(1)\left(\frac{1}{8}\right)\\
&= \frac{1}{8} - \frac{5}{24} - \frac{7}{24} + \frac{1}{8} = -\frac{1}{4}.
\end{aligned}$$

Por lo tanto,

$$\text{Cov}(S,T) = \mathbf{E}(ST) - \mathbf{E}(S)\mathbf{E}(T) = -\frac{1}{4} - \left(-\frac{1}{12}\right)(0) = -\frac{1}{4}.$$

Con todo esto,

$$\rho(S,T) = \frac{\text{Cov}(S,T)}{\sqrt{\text{Var}(S)\,\text{Var}(T)}} = \frac{-\frac{1}{4}}{\sqrt{\left(\frac{107}{144}\right)\times 1}} = \frac{-3}{\sqrt{107}} = -0.29.$$

2. La función de densidad conjunta de las variables aleatorias S y T está definida por

$$f_{XY}(x,y) = \begin{cases} 2, & \text{si } 0 \le x \le 1,\ 0 \le y \le 1,\ x+y \le 1; \\ 0, & \text{caso contrario.} \end{cases}$$

Hallar la correlación entre X y Y.

Solución: Anteriormente determinamos que

$$f_X(x) = 2(1-x) \quad \text{y} \quad f_Y(y) = 2(1-y).$$

Sus esperanzas y varianzas son

$$\mathbf{E}(X) = \mathbf{E}(Y) = \frac{1}{3} \quad \text{y} \quad \text{Var}(X) = \text{Var}(Y) = \frac{1}{18}.$$

También,

$$\mathbf{E}(XY) = \int_0^1 \int_0^{1-x} 2xy\,dy\,dx = \frac{1}{12}.$$

Entonces,

$$\text{Cov}(X,Y) = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y) = \frac{1}{12} - \frac{1}{3} \times \frac{1}{3} = -\frac{1}{36}$$

De manera que

$$\rho(S,T) = \frac{\text{Cov}(S,T)}{\sqrt{\text{Var}(S)\,\text{Var}(T)}} = \frac{-\frac{1}{36}}{\sqrt{\frac{1}{18} \times \frac{1}{18}}} = -\frac{1}{2}.$$

Notemos que estas dos variables no son independientes.

3. Determine la correlación entre las variables aleatorias X y Y, cuya densidad conjunta es

$$f_{X,Y}(x,y) = \frac{1}{2\pi}\frac{1}{\sqrt{1-R^2}}e^{-(x^2+y^2-2Rxy)/(2(1-R^2))}, \quad -\infty < x < \infty, \quad \infty < y < \infty,$$

donde $|R| < 1$ (a esta ley se le denomina *normal bivariante*).

Solución: Calculemos las funciones de densidad de cada una de las variables aleatorias.

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} \frac{1}{2\pi}\frac{1}{\sqrt{1-R^2}}e^{-(x^2+y^2-2Rxy)/(2(1-R^2))}\,dx \\ &= \frac{1}{2\pi}\frac{1}{\sqrt{1-R^2}}e^{-y^2/2}\int_{-\infty}^{\infty} e^{-(x-Ry)^2/(2(1-R^2))}\,dx. \end{aligned}$$

Haciendo $\dfrac{x - Ry}{\sqrt{1-R^2}} = z$ y $dx = \sqrt{1-R^2}dz$, encontramos:

$$f_Y(y) = \frac{1}{2\pi}e^{-y^2/2}\int_{-\infty}^{\infty} e^{-z^2/2}\,dz = \frac{1}{\sqrt{2\pi}}e^{-y^2/2}, \quad -\infty < y < \infty.$$

Análogamente, se determina que $f_X(x) = \dfrac{1}{\sqrt{2\pi}}e^{-x^2/2}, \quad -\infty < x < \infty.$

A partir de ésto, es fácil verificar que $\mathbf{E}(X) = \mathbf{E}(Y) = 0$ y que $\mathrm{Var}(X) = \mathrm{Var}(Y) = 1$.

Encontremos la covarianza:

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{X,Y}(x,y)\,dx\,dy - 0\\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} ye^{-y^2/2}\left(\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi(1-R^2)}}xe^{-(x-Ry)^2/(2(1-R^2))}\,dx\right)dy.\end{aligned}$$

La integral interior es igual a Ry; por lo tanto,

$$\mathrm{Cov}(X,Y) = \frac{R}{\sqrt{2\pi}}\int_{-\infty}^{\infty} y^2e^{-y^2/2}\,dy = R.$$

Con todo esto,

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = \frac{R}{1\cdot 1} = R.$$

5.4. Variables aleatorias multidimensionales

Los conceptos descritos, válidos para variables aleatorias bidimensionales, se pueden generalizar a vectores aleatorios de cualquier dimensión; por lo tanto, solo vamos a exponer las definiciones de manera resumida.

Definición (de variable aleatoria multidimensional) Sobre un mismo espacio muestral Ω están definidas las variables aleatorias X_1, X_2, ..., X_n; entonces, se dice que $Z = (X_1, X_2, \ldots, X_n)$ es un vector aleatorio o una variable aleatoria $n-$dimensional.

Si X_1, X_2, ..., X_n son variables aleatorias discretas, el vector aleatorio Z es discreto y su función de probabilidad es

$$f_Z(x_1,\ldots,x_n) = \mathrm{Pr}(X_1 = x_1, \ldots, X_n = x_n).$$

Si $X_1, \ldots, X_n$ son variables aleatorias continuas, el vector aleatorio Z es continuo y la probabilidad del evento «$(X_1,\ldots,X_n) \in E \subset \mathbf{R}^n$» se calcula por

$$\mathrm{Pr}[(X_1,\ldots,X_n)\in E] = \int\cdots\int_E f_Z(x_1,\ldots,x_n)\,dx_1\cdots dx_n.$$

La función f se denomina *densidad conjunta* de X_1, X_2, ..., X_n.

La función de distribución de la variable aleatoria multivariante Z está definida por

$$F_Z(x_1,\ldots,x_n) = \mathrm{Pr}(X_1 \leq x_1, \ldots, X_n \leq x_n).$$

Las variables aleatorias X_1, X_2, ..., X_n se llaman independientes si

$$F_Z(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n),$$

o equivalentemente,

$$f_Z(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n).$$

Sea g una función definida de $\mathbf{R}^n$ en $\mathbf{R}$, la esperanza matemática de $g(X_1, \ldots, X_n)$, según la ley de Z, se calcula por

$$\mathbf{E}((g(X_1, X_2, \ldots, X_n)) = \sum_{x_1} \cdots \sum_{x_n} g(x_1, \ldots, x_n) f_Z(x_1, \ldots, x_n),$$

cuando Z es discreta, y por

$$\mathbf{E}((g(X_1, X_2, \ldots, X_n)) = \int \cdots \int_{\mathbf{R}^n} g(x_1, \ldots, x_n) f_Z(x_1, \ldots, x_n)\, dx_1 \cdots dx_n,$$

cuando Z es continua.

La covarianza entre dos componentes cualesquiera X_r y X_m de Z es

$$\mathrm{Cov}(X_r, X_m) = \mathbf{E}\left[(X_r - \mathbf{E}(X_r))(X_m - \mathbf{E}(X_m))\right] = \mathbf{E}(X_r X_m) - \mathbf{E}(X_r)\mathbf{E}(X_m),$$

que también se la suele notar como σ_{rm}.

El coeficiente de correlación entre ellas es

$$\rho(X_r, X_m) = \frac{\mathrm{Cov}(X_r, X_m)}{\sqrt{\mathrm{Var}(X_r)\,\mathrm{Var}(X_m)}}.$$

Al vector aleatorio $Z = (X_1, \ldots, X_n)$ se le asocia su matriz de varianza-covarianza (o simplemente matriz de covarianza) definida como

$$H = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \cdots & \mathrm{Var}(X_n) \end{pmatrix}$$

Destaquemos que, por las propiedades de la covarianza, la matriz H es simétrica.

Finalmente, la relación que da la varianza de una suma de variables aleatorias queda como

$$\mathrm{Var}(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \sum_{1 \le i < j \le n} \mathrm{Cov}(X_i, X_j).$$

5.5. Algunas distribuciones multidimensionales importantes

Al igual que en el caso unidimensional, presentamos las distribuciones multidimensionales de mayor importancia; ellas son: la multinomial, la uniforme y la normal bivariante.

5.5.1. Distribución multinomial

Se dice que el vector aleatorio $k-$dimensional $\mathbf{X} = (X_1, \ldots, X_k)$ sigue una distribución multinomial de parámetros $(n; p_1, \ldots, p_k)$, (donde $0 < p_i < 1$, $\sum_{i=1}^{k} p_i = 1$), si

$$\Pr(\mathbf{X} = \mathbf{N}) = \Pr(X_1 = n_1, \ldots, X_k = n_k) = \frac{n!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k},$$

para $\mathbf{N} = (n_1, \ldots, n_k)$, donde $\sum_{i=1}^{k} n_i = n$.

A un vector aleatorio X que sigue una ley multinomial se lo nota $X \sim \mathcal{M}(n; p_1, \ldots, p_k)$.

La esperanza, varianza y covarianza son, respectivamente, iguales a:

$$\begin{aligned} \mathbf{E}(\mathbf{X}) &= n(p_1, \ldots, p_k), \\ \operatorname{Var}(X_i) &= np_i(1 - p_i), \qquad i = 1, \ldots, k, \\ \operatorname{Cov}(X_i, X_j) &= -np_i p_j, \qquad i \neq j. \end{aligned}$$

La distribución multinomial es la generalización multivariante de la distribución binomial. La distribución marginal de cada una de las componentes X_i es binomial de parámetros (n_i, p_i) y cualquier distribución condicionada es también multinomial.

Esta ley de distribución tiene aplicación en el análisis estadístico de datos cualitativos.

Ejemplo. En una empresa operadora de tarjetas de crédito se registró las causas para la renovación de las tarjetas. Se estableció que 60 % es por pérdida, el 25 % por vencimiento y el 15 % por deterioro. Un día se recibieron 28 solicitudes de renovación de tarjetas. Evaluar la probabilidad de que 15 sean por pérdida, 7 por vencimiento y 6 por deterioro.

Solución: Sean:

$$\begin{aligned} X_1 &= \text{número de renovaciones por pérdida}, \\ X_2 &= \text{número de renovaciones por vencimiento}, \\ X_3 &= \text{número de renovaciones por deterioro}. \end{aligned}$$

Se desea evaluar la probabilidad para $n_1 = 15$, $n_2 = 7$, $n_3 = 6$, $(n_1 + n_2 + n_3 = 28)$:

$$\begin{aligned} \Pr(X_1 = 15, X_2 = 7, X_3 = 6) &= \frac{28!}{15!\,7!\,6!}(0.6)^{15}(0.25)^7(0.15)^6 \\ &= 0.021. \end{aligned}$$

5.5.2. Distribución uniforme

Un vector aleatorio $\mathbf{X} = (X_1, \ldots, X_n)$ tiene distribución uniforme en $S = [a_1, b_1] \times \cdots \times [a_n, b_n] \subset \mathbf{R}^n$ si la función de densidad de probabilidad $f(x_1, \ldots, x_n)$ es

$$f(x_1, \ldots, x_n) = \begin{cases} \dfrac{1}{\prod_{i=1}^{n}(b_i - a_i)}, & \text{si } x \in S; \\ 0, & \text{si } x \notin S. \end{cases}$$

Esta distribución es el análogo multivariante de la distribución uniforme. Las distribuciones marginales de las variables aleatorias X_i $(i = 1, \ldots, n)$ son uniformes con densidad

$$f_{X_i}(x) = \begin{cases} \dfrac{1}{b_i - a_i}, & \text{si } x \in [a_i, b_i]; \\ 0, & \text{si } x \notin [a_i, b_i]. \end{cases}$$

5.5.3. Distribución normal bivariante

Un vector aleatorio $Z = (X, Y)$ tiene distribución normal bivariante no degenerada si su función de densidad conjunta es

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right]\right\},$$

donde ρ es el coeficiente de correlación entre X y Y.

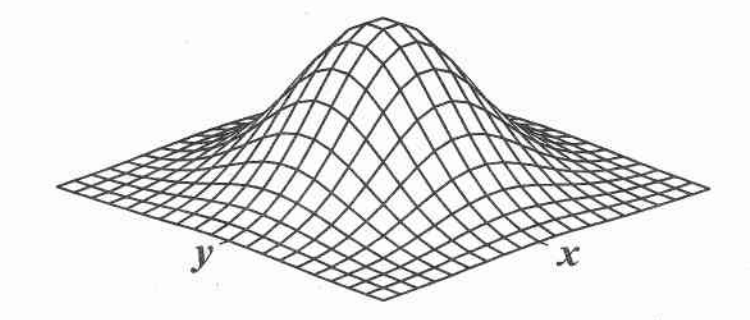


Figura 5.2: Función de densidad de la distribución normal bivariante.

Las leyes marginales de X y de Y son: $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ y $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$.

En un ejemplo de la sección anterior se calculó que el coeficiente de correlación entre X y Y es ρ.

Existe un mayor número de variables aleatorias multidimensionales, pero su tratamiento sale fuera del dominio de esta obra.

5.6. Ejercicios

1. Si la función conjunta de probabilidad de X y Y está dada por

$$f(x,y) = \frac{x+y}{30}, \quad \text{para} \quad x = 0,1,2,3; \quad y = 0,1,2.$$

   Construya una tabla que muestre los valores de la función conjunta de probabilidad de las dos variables aleatorias.

2. Las variables aleatorias S y T tienen la función de probabilidad conjunta que se resume en la siguiente tabla:

	S		
T	0	1	2
0	1/12	1/8	1/24
1	1/4	1/4	1/40
2	1/8	1/20	
3	1/20		

   Encuentre:

a) $\Pr(S=1, T=2)$;
b) $\Pr(S=0, 1 \leq T < 3)$;
c) $\Pr(S+T \leq 2)$;
d) $\Pr(S > T)$;
e) las distribuciones marginales de S y de T.

3. La función conjunta de probabilidad de X y Y está dada por

$$f(x,y) = c(x^2+y^2), \quad \text{para} \quad x=-1,0,1,3; \quad y=-1,2,3.$$

Encuentre:

a) el valor de c;
b) $\Pr(X=0, Y \leq 2)$;
c) $\Pr(X \leq 1, Y > 2)$;
d) $\Pr(X > 2-Y)$;
e) las distribuciones marginales de X y de Y.

4. La distribución conjunta de las variables aleatorias X_1 y X_2 es

	X_1		
X_2	0	1	2
0	p	$p/2$	$p/4$
1	$2p$	p	$p/2$
2	$4p$	$2p$	p

a) Halle el valor de p;
b) Halle las leyes marginales de X_1 y de X_2. ¿Son independientes?;
c) Sean $Y = X_1 \times X_2$, calcule la esperanza de Y.

5. Las variables aleatorias X y Y son independientes entre sí y sus funciones de probabilidad son

i	0	1	2	3	4
$\Pr(X=i)$	0.3	0.2	0.1	0.15	0.25

j	$-\sqrt{3}$	0	$\sqrt{3}$
$\Pr(Y=j)$	0.25	0.67	0.08

Encuentre la función conjunta de probabilidad de (X,Y).

6. Dada la distribución de probabilidad de una variable aleatoria bivariante discreta

	X		
Y	-3	4	10
2	0.15	0.13	0.27
4	0.10	0.30	0.05

a) Halle las leyes de distribución de X y de Y;
b) Calcule el coeficiente de correlación entre X y Y.

7. Dada la distribución de probabilidad de una variable aleatoria bidimensional discreta

	X			
Y	10	20	30	40
0	0.05	0.12	0.08	0.04
1	0.09	0.30	0.11	0.21

a) Determine las leyes de distribución de X y Y;
b) Calcule el coeficiente de correlación entre las variables aleatorias.

8. Se considera la siguiente función de probabilidad conjunta de las variables aleatorias X y Y

$$f_{XY}(x,y) = \begin{cases} k(2x+y), & \text{si } x \in \{0,1,2,3\}, \quad y \in \{1,2,3\}; \\ 0, & \text{caso contrario.} \end{cases}$$

a) Hallar k para que f_{XY} sea función de probabilidad puntual conjunta;

b) Halle las funciones de probabilidad marginal de X y de Y;

c) ¿X y Y son independientes? Justifique la respuesta;

d) Calcule $\Pr(1 \leq X < 3; 2 < Y \leq 3)$ y $\Pr(X+Y < 3)$.

9. Dada la función de distribución de la variable aleatoria bidimensional continua

$$F(x,y) = \begin{cases} \operatorname{sen} x \operatorname{sen} y, & \text{si } \ 0 \leq x \leq \pi/2, \quad 0 \leq y \leq \pi/2; \\ 0, & \text{si } \ x < 0, \quad y < 0. \end{cases}$$

a) Halle la probabilidad de que el punto aleatorio (X,Y) caiga en el rectángulo limitado por las rectas $x = 0$, $x = \frac{\pi}{4}$, $y = \frac{\pi}{6}$, $y = \frac{\pi}{3}$;

b) Determine las funciones de densidad marginal de cada una de las variables aleatorias.

10. Dada la función de densidad conjunta

$$f(x,y) = \frac{1}{\pi} e^{-(x^2+2xy+5y^2)/2}, \quad -\infty < x < \infty, \quad -\infty < y < \infty.$$

Encuentre las funciones de densidad de cada una de las componentes. ¿Son independientes?

11. La variable aleatoria bidimensional (X,Y) tiene función de densidad

$$f(x,y) = \frac{c}{(1+x^2)(16+y^2)}, \quad x,y \in \mathbf{R}.$$

a) Halle la constante c;

b) Determine si X y Y son independientes.

12. La densidad conjunta de X y Y es $f(x;y) = \lambda^3 x e^{-\lambda(x+y)}$ para $x > 0$ y $y > 0$.

a) Halle las densidades marginales y demuestre que X y Y son independientes;

b) Halle $\Pr(X \leq a; Y \leq b)$ para cualesquiera números positivos a y b;

c) Halle $\Pr(X \leq a)$ para $a > 0$.

13. En una investigación sobre la utilización del crédito institucional por parte de los campesinos se registra el número de cultivos que tiene el campesino en su unidad productiva (variable X) y el número de préstamos que ha obtenido en los últimos años (variable Y). En la siguiente tabla se presentan los resultados de la investigación:

	X				
Y	1	2	3	4	5
0	2/50	1/50			
1	3/50	7/50	12/50		
2		5/50	9/50	4/50	3/50
3				3/50	1/50

a) Obtenga las distribuciones marginales de X y de Y;

b) Determine la distribución probabilística de créditos obtenidos, dado que cultiva dos productos;

c) Obtenga la distribución probabilística de productos, ya que no tiene préstamos.

14. Un sociólogo investiga el comportamiento delincuencial de los internos de un penal. La variable X representa el número de veces que ha estado detenido y la variable Y el número de delitos distintos por los que ha sido sentenciado. Sus datos se resumen en la siguiente tabla:

	X				
Y	1	2	3	4	5
1	15/100	9/100	4/100	1/100	1/100
2	5/100	11/100	5/100	2/100	1/100
3	2/100	4/100	7/100	3/100	1/100
4		1/100	3/100	5/100	2/100
5			2/100	4/100	2/100
6			2/100	3/100	1/100
7				2/100	2/100

a) Obtenga las distribuciones marginales de X y de Y;

b) Determine la distribución del número de veces que ha estado detenido, si solo ha cometido un delito.

15. Sea X una variable aleatoria que sigue una ley uniforme sobre $\{1, 2, \ldots, n\}$. Sea Y la variable aleatoria definida por $Y = (X+1)^2$. Calcule la covarianza entre X y Y.

16. Sea X una variable aleatoria que sigue una ley uniforme sobre $\{-1, 0, 1\}$. Calcule el coeficiente de correlación entre X^m y X^n.

17. En una universidad se toma, a los aspirantes, pruebas de ingreso en ciencias y en humanidades. Si X y Y son, respectivamente, las proporciones de respuestas correctas que un estudiante alcanza en las pruebas y su función de densidad conjunta viene dada por

$$f(x,y) = \begin{cases} \dfrac{4x+6y}{5}, & \text{para } 0 \leq x \leq 1; \quad 0 \leq y \leq 1; \\ 0, & \text{caso contrario.} \end{cases}$$

¿Qué porcentaje de estudiantes conseguirán:

a) menos del 40 % de respuestas correctas en cada una de las pruebas?;

b) más del 80 % de respuestas correctas en ciencias y menos de 50 % en humanidades?

18. La cantidad en miligramos de dos componentes contenidos en un producto es una variable aleatoria bivariante, cuya función de densidad viene dada por la expresión

$$f(x,y) = \begin{cases} cxy, & \text{para } 0 \leq x \leq 1; \quad 0 \leq y \leq 1; \\ 0, & \text{caso contrario.} \end{cases}$$

a) Encuentre el valor de la constante c;

b) Halle la ley condicional $f(x|y_0)$;

c) Calcule la probabilidad de que la cantidad del primer componente sea menor que 0.33 miligramos cuando la del segundo es 0.8 miligramos;

d) ¿Son independientes los dos componentes?

19. Si X es la proporción de personas que responden a una encuesta realizada por correo y Y es la proporción de personas que responden a otra encuesta realizada por correo, y la función de densidad conjunta de X y Y está dada por

$$f(x,y) = \begin{cases} \dfrac{2x+8y}{5}, & \text{para } 0 \leq x \leq 1; \quad 0 \leq y \leq 1; \\ 0, & \text{caso contrario.} \end{cases}$$

Encuentre:

a) la función de densidad marginal de X;

b) la probabilidad de que al menos un 30 % de las personas responda a la primera encuesta;

c) la probabilidad de que menos de un 40 % de las personas responda a la primera encuesta y que más de 50 % de las personas responda a la segunda.

20. La vida de uso (en horas) de cierta clase de circuitos integrados es una variable aleatoria con función de densidad

$$f(x) = \begin{cases} \dfrac{20\,000}{(x+100)^3}, & \text{si } x > 0; \\ 0, & \text{caso contrario.} \end{cases}$$

Si tres de estos circuitos operan independientemente, encuentre:

a) la densidad conjunta de X_1, X_2 y X_3, que representan la duración de cada uno de los circuitos;

b) la probabilidad $\Pr(X_1 < 100, X_2 < 100, X_3 \geq 200)$.

21. Sean S y T dos variables aleatorias cuya función de densidad conjunta está dada por

$$f(s,t) = \begin{cases} \dfrac{k}{56}, & \text{si } 0 \leq s \leq 8; \quad 0 \leq t \leq 7; \\ 0, & \text{caso contrario.} \end{cases}$$

a) Encuentre el valor de k;

b) Obtenga las densidades marginales de S y de T;

c) Determine la función de distribución $F(s,t)$.

22. Una función de densidad conjunta está dada por

$$f(x,y,z,t) = \begin{cases} 16xyzt, & \text{si } 0 \leq x \leq 1; \quad 0 \leq y \leq 1; \quad 0 \leq z \leq 1; \quad 0 \leq t \leq 1; \\ 0, & \text{caso contrario.} \end{cases}$$

a) Calcule la probabilidad de que $X \leq \frac{3}{4}$, $Y \geq \frac{1}{2}$, $Z < \frac{1}{2}$ y $T \geq \frac{2}{3}$;

b) Obtenga la densidad marginal de T;

c) ¿Son las variables aleatorias X, Y, Z y T mutuamente independientes?

23. La densidad conjunta de (X,Y) es

$$f(x,y) = \begin{cases} k(6-x-y), & \text{si } 0 \leq x \leq 2; \quad 2 \leq y \leq 4; \\ 0, & \text{caso contrario.} \end{cases}$$

a) Halle el valor de k;

b) Obtenga las densidades marginales de X y de Y;

c) Determine la covarianza entre X y Y.

24. Sea

$$f(x,y) = \begin{cases} cxye^{-(x^2+y^2)}, & x \geq 0; \quad y \geq 0; \\ 0, & \text{en otros casos.} \end{cases}$$

a) Halle el valor de c;

b) Obtenga las densidades marginales de X y de Y;

c) Calcule las esperanzas de X y de Y;

d) ¿Son independientes X y Y?

25. Sean X y Y dos variables aleatorias cuya densidad conjunta es

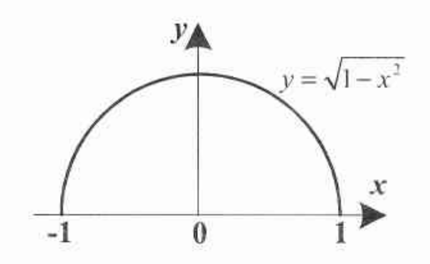
$$f(x,y) = \begin{cases} \frac{4}{5}(6x^2 + y^2)y, & \text{si } 0 \leq x \leq 1; \quad 0 \leq y \leq 1; \\ 0, & \text{caso contrario.} \end{cases}$$

a) Obtenga las distribuciones marginales de X y de Y;

b) Halle la covarianza entre X y Y;

c) Calcule las esperanzas de $X + Y$ y de $X^2 + Y^2$.

26. Sea (X, Y) distribuido uniformemente sobre el semicírculo del diagrama. Entonces, $f(x, y) = \frac{2}{\pi}$ si (x, y) está en el semicírculo.



a) Determine las distribuciones marginales de X y de Y;

b) ¿Son independientes X y Y?

27. Para la distribución bivariante

$$f(x,y) = \begin{cases} \dfrac{c + x + y}{(1+x)^4(1+y)^4}, & \text{si } x \geq 0; \quad y \geq 0; \\ 0, & \text{en otros casos.} \end{cases}$$

a) Halle el valor de c;

b) Obtenga las distribución marginal de X;

c) ¿Son independientes X y Y?

28. Para la siguiente función de densidad:

$$f(x,y) = \begin{cases} \dfrac{k}{(1+x+y)^n}, & \text{si } x \geq 0, \quad y \geq 0; \quad \text{con } n \geq 2, \\ 0, & \text{caso contrario.} \end{cases}$$

a) Determine la constante k;

b) Obtenga la función de distribución.

29. ¿Son independientes las variables aleatorias X y Y, si la función de densidad conjunta es

a) $f(x,y) = 3\dfrac{x^2}{y^3}, \quad 0 \leq x \leq y \leq 1$?;

b) $f(x,y) = \dfrac{6}{(1+x+y)^4}, \quad x \geq 0;\ y \geq 0$?

30. Dada la función de densidad del vector aleatorio (X, Y)

$$f(x,y) = \begin{cases} 6x, & \text{para } 0 \leq x \leq y \leq 1; \\ 0, & \text{caso contrario.} \end{cases}$$

Encuentre las funciones de densidad marginal de X y de Y.

31. Si (X, Y) es un vector aleatorio con función de densidad conjunta $f(x,y) = \frac{6}{5}(x^2+y)$ si $0 < x < 1$ y $0 < y < 1$ (y 0 en otro caso). Calcule la covarianza y el coeficiente de correlación de ambas variables.

32. Dada la función de densidad de (X, Y):

$$f(x,y) = \begin{cases} kxy, & \text{para } 0 \leq x \leq y \leq 1; \\ 0, & \text{caso contrario.} \end{cases}$$

a) Encuentre el valor de k;
b) Calcule las funciones de densidad marginal de X y de Y;
c) Calcule las esperanzas de X y de Y;
d) Calcule $\Pr(X < 0.5 | Y = 0.6)$;
e) ¿Son X y Y independientes?

33. Sea (X, Y) una variable aleatoria bidimensional con función de densidad conjunta

$$f(x,y) = kx^2y, \quad 0 < y < x < 1.$$

Calcule:

a) el valor de la constante k;
b) las funciones de densidad marginal de X y Y. ¿Son independientes?;
c) la covarianza entre X y Y;
d) la función de densidad de $Y|X = \frac{1}{2}$;
e) la esperanza de $Y|X = \frac{1}{2}$.

34. Si (X, Y) está uniformemente distribuido en el triángulo limitado por las rectas $x = 0$, $y = 0$ y $x + y = 2$, encuentre:

a) la función de densidad de (X, Y);
b) las funciones de densidad de X y de Y;
c) la covarianza entre X y Y.

35. La distribución conjunta de las variables aleatorias X y Y es uniforme en el cuadrado con vértices en $(1, 0)$, $(-1, 0)$, $(0, 1)$ y $(0, -1)$.

a) Escriba la función de densidad conjunta de X y Y;

b) Calcule la función de densidad marginal de X y dibújela;

c) Halle $\mathbf{E}(X)$ y $\text{Var}(X)$;

d) Calcule $\mathbf{E}(XY)$;

e) Calcule la covarianza y la correlación entre X y Y.

36. El departamento de señalización del municipio registró el porcentaje de focos que tiene que reemplazar en los semáforos, según el color que ellos iluminan. Se detectó que el 45 % son del color verde, 20 % del color amarillo y 35 % del color rojo. En una muestra aleatoria de 15 focos, ¿cuál es la probabilidad de que:

a) hayan 10 verdes, 1 amarillo y 4 rojos?; b) hayan 5 focos de cada color?

37. Entre las operaciones que se realizan en un banco se ha registrado que a lo largo del tiempo se tienen los siguientes porcentajes de retiros, depósitos y cambios de cheques: 35 %, 40 % y 25 % respectivamente. Un cajero realizó 20 operaciones en una hora. Determine la probabilidad de que se hayan hecho: a) 5 retiros, 10 depósitos y 5 cambios de cheques; b) 10 retiros y 10 depósitos.

38. Según el Registro Civil, la población ecuatoriana entre los 18 y 65 años de edad tiene la siguiente composición: 30 % son solteros, 45 % son casados, 15 % son divorciados y 10 % son viudos. En una oficina laboran 18 personas, ¿cuál es la probabilidad de que hayan:

a) 6 solteros, 6 casados, 3 divorciados y 3 viudos;

b) 7 solteros, 8 casados, 2 divorciados y 1 viudos?

39. De acuerdo a la teoría de la herencia de Mendel, si plantas con semilla amarilla lisa se cruzan con plantas de semilla verde rugosa, se obtiene los siguientes resultados con sus respectivas probabilidades:

Semillas	Probabilidad
Amarilla y lisa	9/16
Amarilla y rugosa	3/16
Verde y lisa	3/16
Verde y rugosa	1/16

¿Cuál es la probabilidad de que entre 9 plantas así obtenidas, 4 sean de semilla amarilla lisa, 2 sean de semilla amarilla rugosa, 3 de semilla verde lisa y ninguna de semilla verde rugosa?

40. Las variables aleatorias X_1, X_2 y X_3 siguen las siguientes leyes de probabilidad: $X_1 \sim \mathcal{N}(10, 1)$, $X_2 \sim \mathcal{N}(20, 1)$ y $X_3 \sim \mathcal{N}(30, 4)$. Se definen

$$Z_1 = X_1 + X_2 - X_3, \quad Z_2 = X_1 + X_2 + X_3, \quad Z_3 = X_1 - X_2 - X_3.$$

Si X_1, X_2, X_3 son independientes calcule la matriz de covarianzas de (Z_1, Z_2, Z_3).

41. Las variables aleatorias X_1, X_2, ..., X_n, Y_1, Y_2, ..., Y_n son independientes. Pongamos

$$S_n = X_1 + X_2 + \cdots + X_n,$$

$$T_n = X_1Y_1 + X_2Y_2 + \cdots + X_nY_n.$$

Halle la covarianza entre S_n y T_n si

$$\mathbf{E}(X_k) = a, \quad \text{Var}(X_k) = \sigma^2,$$

$$\Pr(Y_k = 1) = p, \quad \Pr(Y_k = 0) = q = 1 - p, \quad k = 1, 2, \ldots, n.$$

42. Sea $X \sim \mathcal{N}(1, 1)$, halle la matriz de covarianza de (X, X^2, X^3).

Capítulo 6

Distribuciones de Muestreo

En cierto sentido, la estadística y la probabilidad tratan con problemas inversos: si el objetivo básico de la probabilidad es calcular las probabilidades de eventos complicados que siguen un modelo probabilístico, la estadística trata de clarificar la estructura de modelos probabilístico-estadísticos mediante la observación de varios eventos complicados.
A. N. Shiryayev

La ley de distribución normal es el modelo probabilístico más empleado en la Estadística, debido a la aplicación del Teorema del Límite Central; sin embargo, este resultado no es aplicable en todos los casos. En el presente capítulo examinaremos las leyes de probabilidad que siguen ciertas medidas estadísticas, obtenidas a partir de las muestras, que nos permitirán construir modelos inferenciales sobre los datos.

Las distribuciones de muestreo, constituyen el punto de transición desde la Probabilidad a la Estadística.

6.1. Reseña histórica

La Estadística actual es el resultado de la unión de dos disciplinas que evolucionaron independientemente hasta confluir en el siglo XIX: la primera es el cálculo de las probabilidades, la segunda es la «Estadística» (o ciencia del Estado, del latín *status*), que estudia la descripción de datos, y tiene raíces más antiguas. La integración de ambas líneas de pensamiento dio lugar a una ciencia que estudia cómo obtener conclusiones de la investigación empírica mediante el uso de modelos matemáticos.

Los comienzos de la estadística se pueden hallar en el antiguo Egipto, cuyos faraones recopilaron, hacia el año 3050 antes de Cristo, datos relativos a la población y la riqueza del país. De acuerdo a Heródoto, dicho registro de riqueza y población se hizo con el objetivo de preparar la construcción de las pirámides.

Los chinos, también efectuaron censos hace más de cuarenta siglos. Los griegos realizaron censos periódicamente con fines tributarios, sociales y militares. La investigación histórica revela que se realizaron 69 censos para calcular los impuestos, determinar los derechos de voto y ponderar la potencia guerrera.

Pero fueron los romanos quienes mejor supieron emplear los recursos de la estadística. Cada cinco años realizaban un censo de la población y sus funcionarios públicos tenían la obligación de anotar

nacimientos, defunciones y matrimonios, sin olvidar los recuentos periódicos del ganado y de las riquezas contenidas en las tierras conquistadas.

Durante los mil años siguientes a la caída del imperio Romano se realizaron muy pocas investigaciones estadísticas. El primer intento de aplicar un razonamiento propiamente estadístico, en el sentido actual del término, a datos demográficos es debido a John Graunt, en 1662, quien se planteó el problema de estimar la población inglesa de la época.

Godofredo Achenwall, profesor de la Universidad de Gotinga, acuñó en 1760 la palabra estadística, que extrajo del término italiano *statista* (estadista). Creía, y con sobrada razón, que los datos de la nueva ciencia serían el aliado más eficaz de los gobernantes conscientes.

Durante el siglo XVIII y la mayor parte del siglo XIX, la Estadística evolucionó como ciencia separada del Cálculo de Probabilidades. Una contribución importante al desarrollo de la Estadística es debida a A. Quetelet (1846), quien sostuvo la importancia del cálculo de probabilidades para el estudio de datos humanos. Quetelet demostró que la estatura de los reclutas de un regimiento seguía una ley probabilística, e introdujo el concepto de «hombre medio».

A finales del siglo XIX, Sir Francis Galton ideó el método conocido por correlación, que tenía por objeto medir la influencia relativa de los factores. Sus investigaciones se dirigieron a aplicar métodos cuantitativos en el estudio de la herencia humana. La importancia de Galton radicó no solamente en el nuevo enfoque que introdujo en los problemas de estadística, sino también en su influencia directa sobre W. Weldon, K. Pearson y Edgeworth, entre otros. Además, fundó el primer departamento de Estadística.

Pero, talvez quien más ha influido en el desarrollo de la Estadística moderna es R. A. Fisher (1890 - 1962). Fisher se interesó primero por la eugenesia, lo que le condujo, siguiendo los pasos de Galton, a la investigación estadística. Sus trabajos culminaron con la publicación del libro *Statistical Methods for Research Workers*. En esta obra aparece el cuerpo metodológico básico de la Estadística actual.

A partir de 1950 se puede considerar que comienza la época moderna de la Estadística. Un aspecto diferencial respecto a los periodos anteriores es la aparición de las computadoras, que revolucionaron la metodología estadística y abren enormes posibilidades para la construcción de modelos complejos.

En la actualidad, la Estadística es una disciplina que actúa como puente entre los modelos matemáticos y los fenómenos reales. Un modelo es una abstracción simplificada de una realidad más compleja y siempre existirá discrepancia entre lo observado y lo previsto por el modelo. La Estadística proporciona una metodología para evaluar y juzgar estas discrepancias entre la realidad y la teoría.

6.2. Definiciones básicas

A continuación damos varias definiciones de interés, que permitirán entender la terminología que emplearemos. Algunas definiciones ya se dio con anterioridad, pero las repetimos para una mayor claridad de los conceptos.

Definición (de población) Una población (o universo) es una colección completa de personas, animales, plantas o cosas de las cuales se desea recolectar datos. Es el grupo entero al que queremos describir o del que deseamos sacar conclusiones.

La población debe tener características medibles o contables, de naturaleza cuantitativa o cualitativa. A la característica medible se denomina *variable estadística* y a los valores que toma se los llama *observaciones*.

Definición (de muestra) Es un grupo de unidades seleccionadas de un grupo mayor (la población). Por el estudio de la muestra se espera obtener conclusiones sobre la población.

Definición (de parámetro) Un parámetro es un valor, usualmente desconocido (y que por lo tanto tiene que ser estimado), usado para representar cierta característica de la población.

Entre otros, los parámetros poblacionales son:

- La media, μ;
- El total, τ;
- La varianza, σ^2;
- La desviación estándar, σ;
- La proporción, π o p.

Definición (de estadístico) Un estadístico es una cantidad que se calcula a partir de una muestra de datos. Se los emplea para dar información sobre los valores desconocidos correspondientes a la población.

Por ejemplo, el promedio de los datos de una muestra, se usa para dar información sobre la media de la población, de la cual se extrajo la muestra.

Generalmente, a los estadísticos se les asigna letras latinas (por ejemplo, m y s); en cambio, a los parámetros poblacionales se les asigna letras griegas (por ejemplo, μ y σ).

Dentro de una población, un parámetro es un valor fijo que no varía; mientras que es posible extraer más de una muestra de la misma población y el valor de un estadístico variará de muestra a muestra; por ello, un estadístico es una variable aleatoria que sigue una ley de probabilidad.

Los estadísticos más importantes y sus valores, calculados a partir de una muestra de tamaño n, son:

- La media muestral o promedio, $\overline{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$;
- El total muestral, $\widehat{\tau} = N\overline{x}$, donde N es el tamaño de la población;
- La varianza muestral, $s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$;
- La desviación estándar muestral, $s = \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$;
- La proporción muestral, $\widehat{p} = \dfrac{y}{n}$, donde y es el número de éxitos entre n intentos.

6.3. Distribuciones de muestreo

Si decimos que un estadístico es una variable aleatoria, entonces tendrá una ley de probabilidad asociada.

Definición (de distribución de muestreo) A la ley de probabilidad que sigue un estadístico se le denomina distribución de muestreo.

La derivación de la distribución de muestreo es el primer paso en la realización de inferencias sobre el valor del parámetro asociado al estadístico que se estudia.

6.3.1. Distribución de muestreo de la media

Supongamos que se obtiene una muestra $X_1, X_2, \ldots, X_n$ de una población que tiene media μ y varianza σ^2. A partir de la muestra calculamos el promedio, $\overline{X}$. Entonces, se cumple que:

1. $\mathbf{E}(\overline{X}) = \mu$;
2. $\mathrm{Var}(\overline{X}) = \dfrac{\sigma^2}{n}$;
3. $\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ sigue aproximadamente una ley normal estándar (por el Teorema del Límite Central). Es decir,
$$\Pr(\overline{X} \leq t) \approx \Pr\left(Z \leq \frac{t-\mu}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{t-\mu}{\sigma/\sqrt{n}}\right),$$
donde Z es una variable aleatoria normal estándar.

Téngase en cuenta que, para la mayoría de aplicaciones, ya se obtiene una buena aproximación con un tamaño de muestra de $n = 25$.

Observación. La desviación estándar de la media muestral se denomina *error estándar* y se le nota $\sigma_{\overline{x}}$:
$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}.$$

Ejemplos

1. Supongamos que se selecciona una muestra de $n = 36$ observaciones de una población con $\mu = 7$ y $\sigma = 0.9$.

   a) Hallar la probabilidad aproximada de que la media muestral sea menor que 6.9.
   b) Obtener la probabilidad aproximada de que el promedio $\overline{X}$ sea mayor que 6.82.
   c) Hallar la probabilidad para que $\overline{X}$ esté en el intervalo (6.8; 7.29).

   Solución: La distribución de la media muestral $\overline{X}$ sigue una ley normal con media $\mu = 7$ y varianza $\dfrac{\sigma^2}{n} = \dfrac{(0.9)^2}{36}$; es decir, $\dfrac{\sigma}{\sqrt{n}} = \dfrac{0.9}{\sqrt{36}} = 0.15$.

a) Así,

$$\begin{aligned}\Pr(\overline{X} < 6.9) &\approx \Pr\left(Z < \frac{6.9-7}{0.15}\right) = \Phi(-0.67)\\ &= 0.2515.\end{aligned}$$

b) Tenemos que

$$\begin{aligned}\Pr(\overline{X} > 6.82) &= 1 - \Pr(\overline{X} \leq 6.82) \approx 1 - \Pr\left(Z \leq \frac{6.82-7}{0.15}\right)\\ &= 1 - \Phi(-1.2) = 1 - 0.1151\\ &= 0.8849.\end{aligned}$$

c) La probabilidad de que la media muestral esté entre 6.8 y 7.29 es

$$\begin{aligned}\Pr(6.8 < \overline{X} < 7.29) &= \Pr\left(\frac{6.8-7}{0.15} < Z < \frac{7.29-7}{0.15}\right)\\ &= \Phi(1.93) - \Phi(-1.33) = 0.9732 - 0.0918\\ &= 0.8814.\end{aligned}$$

2. El número de clientes que ocupan un cajero automático, en un lapso de 5 minutos, es una variable aleatoria distribuida según la siguiente ley de probabilidad:

k	0	1	2	3	4	5
p_k	1/12	2/12	3/12	3/12	2/12	1/12

a) Halle la media y la varianza de la variable aleatoria; b) Se escogieron 46 muestras, de 5 minutos de duración, y se contó el número clientes que utilizaron el cajero. ¿Cuál es la probabilidad de que el promedio de clientes esté entre 2.2 y 3?

Solución:

a) La media y la varianza son:

$$\begin{aligned}\mu &= \sum k p_k = 0 \times \frac{1}{12} + 1 \times \frac{2}{12} + 2 \times \frac{3}{12} + 3 \times \frac{3}{12} + 4 \times \frac{2}{12} + 5 \times \frac{1}{12}\\ &= \frac{5}{2},\\ \sigma^2 &= \sum k^2 p_k - \mu^2 = 0 \times \frac{1}{12} + 1 \times \frac{2}{12} + 4 \times \frac{3}{12} + 9 \times \frac{3}{12} + 16 \times \frac{2}{12} + 25 \times \frac{1}{12} - \left(\frac{5}{2}\right)^2\\ &= \frac{23}{12}.\end{aligned}$$

b) Sea $\overline{X}$ la media de clientes que ocupan el cajero en muestras de tamaño 46. Se tiene que

$$\mu_{\overline{x}} = \mu = \frac{5}{2} \quad \text{y} \quad \sigma^2_{\overline{x}} = \frac{\sigma^2}{n} = \frac{\frac{23}{12}}{46} = \frac{1}{24}.$$

Entonces, $\overline{X} \sim \mathcal{N}\left(\frac{5}{2}, \frac{1}{24}\right)$; por lo tanto,

$$\begin{aligned}\Pr\left(2.2 \leq \overline{X} \leq 3\right) &= \Pr\left(\frac{2.2 - \frac{5}{2}}{\sqrt{\frac{1}{24}}} \leq Z \leq \frac{3 - \frac{5}{2}}{\sqrt{\frac{1}{24}}}\right)\\ &= \Phi(2.45) - \Phi(-1.47) = 0.9929 - 0.0708\\ &= 0.9221.\end{aligned}$$

3. En una planta pasteurizadora se ha observado que la máquina que llena las fundas de leche, envasa el líquido con una media μ y una desviación estándar de $\sigma = 20\,\text{cm}^3$. Si un día se llevan a cabo 25 mediciones de la cantidad de leche en cada funda. a) Calcular la probabilidad de que el promedio medido difiera a lo mucho en $8\,\text{cm}^3$ de la media teórica que debe tener el volumen de leche envasado; b) ¿cuántas mediciones deben realizarse para que $\overline{x}$ difiera de μ en menos de $8\,\text{cm}^3$, con una probabilidad de 0.99?

Solución: Como $n = 25$, se puede asumir que la distribución de $\overline{X}$ es aproximadamente normal.

a) Entonces,

$$\begin{aligned} \Pr(|\overline{X} - \mu| \leq 8) &= \Pr(-8 \leq \overline{X} - \mu \leq 8) \\ &= \Pr\left(-\frac{8}{20/\sqrt{25}} \leq Z \leq \frac{8}{20/\sqrt{25}}\right) \\ &= \Pr(-2 \leq Z \leq 2), \end{aligned}$$

donde $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ sigue una distribución normal estándar. La probabilidad buscada es

$$\begin{aligned} \Pr(-2 \leq Z \leq 2) &= \Phi(2) - \Phi(-2) \\ &= 0.9772 - 0.0228 = 0.9544. \end{aligned}$$

b) Se tiene que

$$\Pr(|\overline{X} - \mu| \leq 8) = \Pr(-8 \leq \overline{X} - \mu \leq 8) = 0.99.$$

Como $\sigma = 20$,

$$\Pr\left(-\frac{8\sqrt{n}}{20} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{8\sqrt{n}}{20}\right) = \Pr(-0.4\sqrt{n} \leq Z \leq 0.4\sqrt{n}) = 0.99.$$

Mediante la tabla de la ley normal se encuentra que

$$\Pr(-2.57 \leq Z \leq 2.57) = 0.99,$$

por lo que se deduce que $0.4\sqrt{n} = 2.57$, o sea

$$n = \left(\frac{2.57}{0.4}\right)^2 = 40.96.$$

Se necesitan al menos 41 mediciones para que el promedio de la muestra esté a $8\,\text{cm}^3$ de la media poblacional con probabilidad del 99 %.

6.3.2. Distribución de muestreo de la proporción

Supóngase que tenemos una muestra aleatoria $X_1, X_2, \ldots, X_n$, proveniente de una población que sigue una ley de Bernoulli, $\mathbf{Ber}(p)$. Definimos

$$Y = \sum_{i=1}^{n} X_i,$$

donde $X_i = 1$ con probabilidad p y $X_i = 0$ con probabilidad $q = 1 - p$, $i = 1, 2, \ldots, n$. Entonces, Y cuenta el número de éxitos en n intentos. La proporción de «éxitos» en la muestra es

$$\widehat{p} = \frac{Y}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

La variable aleatoria Y tiene distribución binomial de parámetros (n, p). Por lo que, $\mu_Y = np$ y $\sigma_Y^2 = npq$, y se cumple que:

1. $\mathbf{E}(\widehat{p}) = \frac{1}{n}\mathbf{E}(Y) = p$;

2. $\mathrm{Var}(\widehat{p}) = \frac{1}{n^2}\,\mathrm{Var}(Y) = \frac{pq}{n}$;

3. $\frac{\widehat{p}-p}{\sqrt{pq/n}}$ sigue aproximadamente una ley normal estándar (por el Teorema del Límite Central).
   Es decir,
$$\Pr(\widehat{p} \leq t) \approx \Pr\left(Z \leq \frac{t-p}{\sqrt{pq/n}}\right) = \Phi\left(\frac{t-p}{\sqrt{pq/n}}\right),$$
   donde Z es una variable aleatoria normal estándar.

Ejemplos

1. En un proceso de producción el 20 % de los productos tienen algún defecto. Se selecciona una muestra de 100 unidades y se cuenta el número de artículos defectuosos. Determinar la probabilidad de que la proporción de defectuosos se encuentre entre el 15 % y el 29 %.
   Solución: Tenemos que $p = 0.2$ y $\frac{pq}{n} = \frac{0.2 \times 0.8}{100} = \frac{0.16}{100}$.
   Se desea evaluar la probabilidad $\Pr(0.15 \leq \widehat{p} \leq 0.29)$; ésto es,
$$\begin{aligned}\Pr(0.15 \leq \widehat{p} \leq 0.29) &= \Pr(\widehat{p} \leq 0.29) - \Pr(\widehat{p} \leq 0.15)\\ &\approx \Phi\left(\frac{0.29-0.2}{\sqrt{0.16/100}}\right) - \Phi\left(\frac{0.15-0.2}{\sqrt{0.16/100}}\right)\\ &= \Phi(2.25) - \Phi(-1.25) = 0.9878 - 0.1057\\ &= 0.8821.\end{aligned}$$

2. En una investigación por muestreo interesaba saber el nivel de sintonía, en los hogares, de un partido de fútbol. Se realizó una encuesta en Quito a 213 hogares y se encontró que el 53 % de los hogares habían visto el mencionado partido. Supongamos que la proporción p de hogares en los que se vio el partido fue realmente igual a 0.5. ¿Cuál es la probabilidad de observar una proporción muestral $\widehat{p}$ igual o mayor que la observada 0.53?
   Solución: Tenemos $n = 213$ y $p = 0.5$; por lo tanto, $\frac{pq}{n} = \frac{0.5 \times 0.5}{213} = 0.001174$.
   Por lo que,
$$\begin{aligned}\Pr(\widehat{p} \geq 0.53) &= 1 - \Pr(\widehat{p} < 0.53) = 1 - \Pr\left(\frac{\widehat{p}-p}{\sqrt{pq/n}} < \frac{0.53-0.5}{\sqrt{0.001174}}\right)\\ &\approx 1 - \Phi(0.88) = 1 - 0.8105\\ &= 0.1895.\end{aligned}$$
   Es decir, con una muestra de 213 hogares la probabilidad de que la proporción observada sea mayor o igual a 0.53 es del 18.95 %.

3. Hay que tomar una muestra aleatoria para estimar la proporción de artículos defectuosos p de un proceso de producción. a) Establecer el tamaño mínimo de la muestra de modo que la proporción observada difiera de la proporción verdadera en menos de 0.1, con una probabilidad de al menos el 95 %; b) Realizar el inciso anterior si se conoce que la proporción de artículos defectuosos es menor que el 12 %.
   Solución:

a) Deseamos determinar el tamaño mínimo de la muestra de tal modo que $\Pr(|p-\widehat{p}| \leq 0.1) \geq 0.95$:

$$\begin{aligned}
\Pr(|p-\widehat{p}| \leq 0.1) &= \Pr(-0.1 \leq p-\widehat{p} \leq 0.1) \\
&= \Pr\left(\frac{-0.1}{\sqrt{pq/n}} \leq Z \leq \frac{0.1}{\sqrt{pq/n}}\right) \\
&= \Phi\left(\frac{0.1\sqrt{n}}{\sqrt{pq}}\right) - \Phi\left(-\frac{0.1\sqrt{n}}{\sqrt{pq}}\right) \geq 0.95.
\end{aligned}$$

Igualando con los valores de la distribución normal y haciendo uso de su propiedad de simetría,

$$\begin{aligned}
\frac{0.1\sqrt{n}}{\sqrt{pq}} &\geq 1.96 \\
n &\geq (19.6)^2 pq,
\end{aligned}$$

En esta última expresión, el valor de n depende de p, que a su vez es desconocido. Para asegurarnos de tener un tamaño de muestra confiable, se maximiza el producto pq. Tal producto es máximo cuando $p = q = 0.5$; entonces,

$$n \geq (19.6)^2(0.5)(0.5) = 96.$$

La muestra tiene que ser de al menos 96 artículos.

b) Antes habíamos determinado que $n \geq (19.6)^2 pq$.
Como $p < 0.12$, se tiene que $pq > 0.12 \times 0.88 = 0.1056$; entonces,

$$n \geq (19.6)^2(0.1056) = 40.57.$$

El tamaño de la muestra es de al menos 41 artículos.

6.3.3. Distribución de muestreo de la varianza

En el Teorema del Límite Central la convergencia de la ley de $\overline{x}$ está asegurada y es independiente de la ley que siguen las observaciones. Para otros estadísticos, se necesitan hipótesis adicionales para asegurar su convergencia hacia una ley de distribución, este es el caso de la varianza muestral.

La ley de distribución χ^2

Sean $X_1, X_2, \ldots, X_n$, n variables aleatorias independientes que siguen una distribución normal estándar, la variable aleatoria definida por $T = \sum_{i=1}^{n} X_i^2$ tiene una distribución χ^2 (ji-cuadrado) con n grados de libertad (g.l.), denotada $\chi^2(n)$.

Su función de densidad es

$$f(x) = \begin{cases} \dfrac{x^{(n-2)/2} e^{-x/2}}{2^{n/2}\Gamma\left(\dfrac{n}{2}\right)}, & \text{si } x \geq 0; \\ 0, & \text{si } x < 0. \end{cases}$$

La ley de distribución χ^2 fue dada en primer lugar, aparentemente, por Abbe (1863) y redescubierta por Helmert (1875) y K. Pearson (1900).

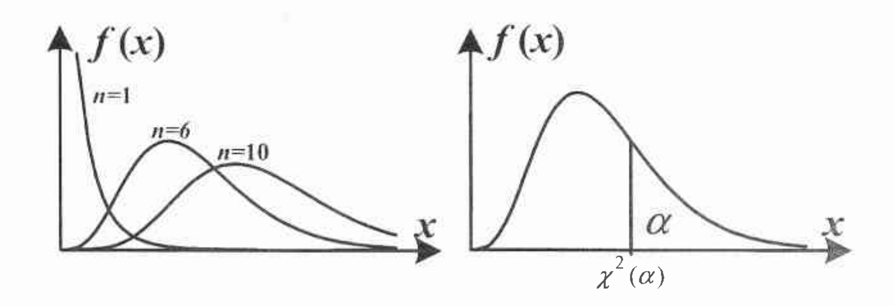


Figura 6.1: Funciones de densidad de la ley χ^2.

Esta distribución está definida para valores mayores que cero y viene tabulada. La Tabla 3 del Apéndice contiene los valores χ^2_α que cortan un área α en el extremo derecho de la distribución (Figura 6.1).

La lectura de la tabla se realiza de la siguiente manera:

1. Escoja el número n de grados de libertad (margen vertical).
2. Considere la probabilidad α en el extremo derecho de la distribución (margen horizontal).
3. Proceda a localizar el valor $\chi^2_\alpha(n)$ en el cuerpo de la tabla, donde se cruzan las líneas vertical y horizontal antes encontradas.

Ejemplo. Se desea conocer el valor de la ley χ^2 a 4 g.l. para el cual el área en el extremo superior es igual a 0.025.

Solución: Se busca $\chi^2_{0.025}(4) = 11.14$. Esto quiere decir que el área a la derecha del valor $t = 11.14$ de la ley χ^2 con 4 g.l. es igual a 0.025: $\Pr\left(\chi^2 > 11.14\right) = 0.025$.

La ley de distribución de s^2

Supongamos que se obtiene una muestra $X_1, X_2, \ldots, X_n$ de una población que sigue una ley normal $\mathcal{N}(\mu, \sigma^2)$. A partir de la muestra calculamos la varianza muestral, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$; entonces, se cumple que:

1. $\mathbf{E}\left(s^2\right) = \sigma^2$;
2. $\operatorname{Var}\left(s^2\right) = \dfrac{2\sigma^4}{n-1}$;
3. $\dfrac{(n-1)s^2}{\sigma^2}$ sigue una ley $\chi^2(n-1)$.

Ejemplo. Un jugador profesional de dardos decide tratar de mejorar su técnica de lanzamiento y va a estudiar la varianza de las distancias al centro del blanco a las que cae el dardo. Para una cierta técnica de lanzamiento se sabe que esas distancias tienen una distribución normal cuya desviación estándar es 4 cm. Realiza 30 lanzamientos y calcula la varianza de las distancias entre el sitio de impacto del dardo y el centro del blanco. a) Calcular la probabilidad de que la desviación estándar de los lanzamientos sea mayor a 3 cm.; b) Hallar $\Pr(10 \leq s^2 \leq 27)$, aproximadamente; c) Calcular la media y la varianza de s^2.

Solución: Sea $U = \dfrac{(n-1)s^2}{\sigma^2}$, $U \sim \chi^2(29)$ para $n = 30$.

a) Se quiere buscar $\Pr(s > 3)$; es decir, $\Pr(s^2 > 9)$:

$$\Pr(s^2 > 9) = \Pr\left(\frac{n-1}{\sigma^2}s^2 > \frac{29}{16}9\right) = \Pr(U > 16.312).$$

En la tabla de la ley χ^2 para 29 g.l., se tiene que $\Pr(U > 16.05) = 0.975$ y $\Pr(U > 17.71) = 0.95$. Por lo tanto, $\Pr(s^2 > 9)$ es algo menor que 0.975.

b) Calculemos $\Pr(10 \leq s^2 \leq 27)$:

$$\begin{aligned}\Pr(10 \leq s^2 \leq 27) &= \Pr\left(\frac{29}{16}10 \leq \frac{n-1}{\sigma^2}s^2 \leq \frac{29}{16}27\right) = \Pr(18.13 \leq U \leq 48.94)\\ &= \Pr(U \leq 48.94) - \Pr(U \leq 18.13)\\ &= [1 - \Pr(U > 48.94)] - [1 - \Pr(U > 18.13)]\\ &= \Pr(U > 18.13) - \Pr(U > 48.94).\end{aligned}$$

De acuerdo a la tabla χ^2, para 29 g.l. se tiene:

- $\Pr(U > 17.71) = 0.95$ y $\Pr(U > 18.85) = 0.925$.
- $\Pr(U > 45.72) = 0.025$ y $\Pr(U > 49.59) = 0.01$.

Entonces, $\Pr(10 \leq s^2 \leq 27) \approx 0.95 - 0.01 = 0.94$.

Observemos que, por la forma de lectura de la tabla, tuvimos que realizar una transformación para encontrar $\Pr(18.13 \leq U \leq 48.94)$ y su valor se determina de forma aproximada.

c) Sabemos que $\mathbf{E}(s^2) = \sigma^2 = 16$ y que $\mathrm{Var}(s^2) = \dfrac{2\sigma^4}{n-1} = \dfrac{2(16)^2}{29} = 17.66$.

6.4. Otras distribuciones de muestreo

En esta sección presentaremos distintas distribuciones de muestreo, que se presentan cuando tratamos con transformaciones adecuadas de los estadísticos. Estas transformaciones son necesarias para obtener leyes de probabilidad que permitan trabajar adecuadamente.

6.4.1. Distribución de muestreo de la media, cuando la varianza es desconocida

Se puede suponer que –debido al Teorema del Límite Central–, siempre se tiene la convergencia hacia la ley normal; pero ésto no siempre sucede; como, por ejemplo, cuando la varianza poblacional es desconocida.

La ley de distribución t de Student

En 1908, W. S. Gosset, escribiendo con el nombre de Student, publicó en la revista *Biometrika* su deducción de la distribución t e incluyó tablas de probabilidad acumulada de la ley.

El gráfico de la función de densidad de la ley t tiene una forma parecida al de la ley normal, simétrica respecto a 0 y se extiende a lo largo del eje real.

Los valores de probabilidad que toma vienen tabulados. La Tabla 2 del Apéndice contiene los valores de t_α que cortan un área igual a α en el extremo derecho de la distribución (Figura 6.2).

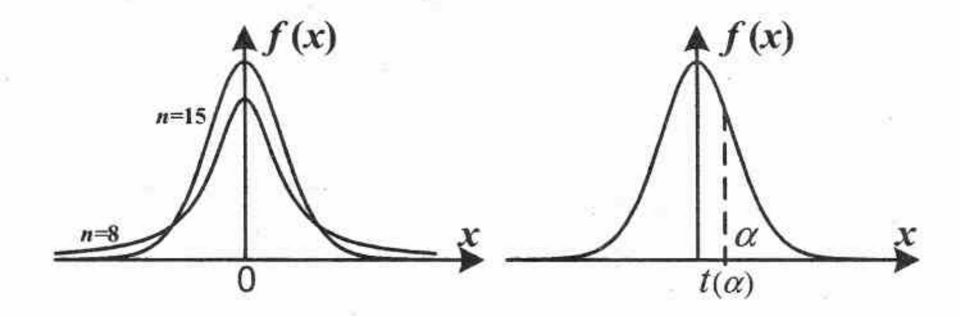


Figura 6.2: Funcion de densidad de la ley t.

Los valores tabulados dependen de los grados de libertad, porque la ley de probabilidad t cambia si n varía. Cuando n aumenta, la distribución t se aproxima a la normal estándar.

La lectura de la tabla se realiza de la siguiente manera:

1. Escoja el número n de grados de libertad (margen vertical).
2. Considere la probabilidad α en el extremo derecho de la distribución (margen horizontal).
3. Proceda a localizar el valor $t_\alpha(n)$ en el cuerpo de la tabla, donde se cruzan las líneas vertical y horizontal antes encontradas.

Ejemplo. Encontrar el valor de la ley t a 6 g.l. para el cual el área en el extremo superior es igual a 0.025.

Solución: Se busca $t_{0.025}(6) = 2.447$. Esto quiere decir que el área a la derecha del valor $t = 2.447$ de la ley t es igual a 0.025: $\Pr(T > 2.447) = 0.025$.

La ley de distribución de $\overline{X}$

Supongamos que se obtiene una muestra X_1, X_2, ..., X_n de una población que sigue una ley normal $\mathcal{N}(\mu, \sigma^2)$, donde σ^2 es desconocida. Entonces, se cumple que la variable aleatoria $T = \dfrac{\overline{X} - \mu}{s/\sqrt{n}}$ sigue una ley t de Student con $(n - 1)$ grados de libertad. Es decir,

$$\Pr(\overline{X} \leq t) = \Pr\left(T \leq \frac{t - \mu}{s/\sqrt{n}}\right).$$

Ejemplo. Un fabricante de cigarrillos asegura que el contenido medio de nicotina, en una de sus marcas, es de 0.6 mg por cigarrillo. Una organización independiente mide el contenido de nicotina de 16 cigarrillos de esta marca y encuentra que el promedio y la desviación estándar son de 0.744 y 0.175 mg de nicotina, respectivamente. Si se supone que la cantidad de nicotina de estos cigarrillos es una variable aleatoria normal, ¿qué tan probable es el resultado obtenido por la organización independiente?

Solución: Se tiene que $\mu = 0.6$, $s = 0.175$, $n = 16$. Encontremos la probabilidad de hallar un valor promedio igual o superior a 0.744.

$$\Pr(\overline{X} \geq 0.744) = \Pr\left(T \geq \frac{0.744 - 0.6}{0.175/\sqrt{16}}\right) = \Pr(T \geq 3.29).$$

De la lectura de la tabla de la ley t con 15 g.l., resulta que $\Pr(T \geq 3.29) = 0.0025$. De manera que el dato proporcionado es muy poco probable.

6.4.2. Distribución de muestreo de la diferencia de dos medias, con varianzas conocidas

Supongamos que se dispone de dos poblaciones que tienen medias μ_1 y μ_2 y varianzas σ_1^2 y σ_2^2, respectivamente. Sean $\overline{X}_1$ y $\overline{X}_2$ las media muestrales de dos muestras aleatorias independientes de tamaños n_1 y n_2, seleccionadas respectivamente de las poblaciones 1 y 2. Entonces, $\overline{X}_1 - \overline{X}_2$ cumple que:

1. $\mu_{\overline{x}_1-\overline{x}_2} = \mathbf{E}(\overline{X}_1 - \overline{X}_2) = \mu_1 - \mu_2$;

2. $\sigma^2_{\overline{x}_1-\overline{x}_2} = \mathrm{Var}(\overline{X}_1 - \overline{X}_2) = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$;

3. Para n_1 y n_2 suficientemente grandes, la variable aleatoria

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

sigue aproximadamente una ley normal estándar. Es decir,

$$\Pr(\overline{X}_1 - \overline{X}_2 \leq t) \approx \Pr\left(Z \leq \frac{t - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}\right) = \Phi\left(\frac{t - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}\right).$$

Para la mayoría de aplicaciones, ya se obtiene una buena aproximación si $n_1 \geq 25$ y $n_2 \geq 25$.

Ejemplo. Una marca de automóviles tiene dos plantas que ensamblan el mismo modelo de autos. El rendimiento medio de estos automóviles debe ser de 40 km por galón, con una desviación estándar de 5 km por galón. La empresa tiene la política de regularmente comparar los rendimientos de los carros ensamblados, escogiendo muestras aleatorias en las dos plantas. Se tomaron sendas muestras de tamaño 30 y se controló el consumo promedio de cada una. Hallar la probabilidad de que la diferencia entre los promedios sea menor a 2 km por galón.

Solución: Se tiene que $\mu_1 = \mu_2 = 40$, $\sigma_1 = \sigma_2 = 5$, $n_1 = n_2 = 30$. Por tanto,

$$\begin{aligned}
\Pr(|\overline{X}_1 - \overline{X}_2| \leq 2) &\approx \Pr\left(-\frac{2-(40-40)}{\sqrt{\dfrac{5^2}{30}+\dfrac{5^2}{30}}} \leq Z \leq \frac{2-(40-40)}{\sqrt{\dfrac{5^2}{30}+\dfrac{5^2}{30}}}\right) \\
&= \Phi(1.55) - \Phi(-1.55) = 0.9394 - 0.0606 \\
&= 0.8788.
\end{aligned}$$

6.4.3. Distribución de muestreo de la diferencia de dos medias, con varianzas desconocidas

Supongamos que se dispone de dos poblaciones que siguen una ley normal: la población 1 sigue una ley $\mathcal{N}(\mu_1, \sigma_1^2)$ y la población 2 sigue una ley $\mathcal{N}(\mu_2, \sigma_2^2)$. Sean $\overline{X}_1$ y $\overline{X}_2$ las media muestrales de dos muestras aleatorias independientes de tamaños n_1 y n_2, seleccionadas respectivamente de las poblaciones 1 y 2.

Caso 1. Las varianzas poblacionales son iguales: $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

La variable aleatoria

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

sigue una ley t con $(n_1 + n_2 - 2)$ g.l., donde $s^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$.

Caso 2. Las varianzas poblacionales son diferentes: $\sigma_1^2 \neq \sigma_2^2$.

La variable aleatoria

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

sigue una ley t con g g.l., donde

$$g = \frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \dfrac{\left(\dfrac{s_2^2}{n_2}\right)^2}{n_2 - 1}}.$$

Cuando g no es un número natural, se redondea al entero más cercano.

6.4.4. Distribución de muestreo de la diferencia de dos proporciones

Supongamos que se dispone de dos poblaciones independientes, que siguen distribuciones de Bernoulli de parámetros p_1 y p_2, respectivamente. De la primera población escogemos una muestra de tamaño n_1: $X_1, X_2, \ldots, X_{n_1}$ y de la segunda población escogemos una muestra de tamaño n_2: $Y_1, Y_2, \ldots, Y_{n_2}$.

Construimos las proporciones muestrales de «éxitos»:

$$\widehat{p}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i \quad \text{y} \quad \widehat{p}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i.$$

Entonces, $\widehat{p}_1 - \widehat{p}_2$ cumple que:

1. $\mu_{\widehat{p}_1 - \widehat{p}_2} = \mathbf{E}(\widehat{p}_1 - \widehat{p}_2) = p_1 - p_2$;
2. $\sigma^2_{\widehat{p}_1 - \widehat{p}_2} = \mathrm{Var}(\widehat{p}_1 - \widehat{p}_2) = \dfrac{p_1(1 - p_1)}{n_1} + \dfrac{p_2(1 - p_2)}{n_2}$;
3. Para n_1 y n_2 suficientemente grandes, la variable aleatoria

$$Z = \frac{(\widehat{p}_1 - \widehat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1 - p_1)}{n_1} + \dfrac{p_2(1 - p_2)}{n_2}}}$$

sigue aproximadamente una ley normal estándar (por el Teorema del Límite Central). Es decir,

$$\Pr(\widehat{p}_1 - \widehat{p}_2 \leq t) \approx \Pr\left(Z \leq \frac{t-(p_1-p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}}\right) = \Phi\left(\frac{t-(p_1-p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}}\right).$$

Ejemplo. Una firma especializada en sondeos políticos afirma que el 30 % de las mujeres y el 20 % de los hombres están a favor de la reelección del actual alcalde. Si se hace un sondeo aleatorio a 150 personas de cada sexo, ¿con qué probabilidad la diferencia entre las proporciones muestrales de las mujeres y de los hombres es, en valor absoluto, menor a 0.19?

Solución: Tenemos que $p_h = 0.3$ y $p_m = 0.2$, $n_h = n_m = 150$. Además,

$$\sigma^2_{\widehat{p}_1-\widehat{p}_2} = \frac{0.3(1-0.3)}{150} + \frac{0.2(1-0.2)}{150} = 0.00247.$$

Buscamos la probabilidad $\Pr(|\widehat{p}_h - \widehat{p}_m| < 0.19)$:

$$\begin{aligned}
\Pr(|\widehat{p}_h - \widehat{p}_m| < 0.19) &= \Pr(-0.19 < \widehat{p}_h - \widehat{p}_m < 0.19) \\
&= \Phi\left(\frac{0.19-(0.3-0.2)}{\sqrt{0.00247}}\right) - \Phi\left(-\frac{0.19-(0.3-0.2)}{\sqrt{0.00247}}\right) \\
&= \Phi(1.81) - \Phi(-1.81) = 0.9648 - 0.0352 \\
&= 0.9296.
\end{aligned}$$

6.4.5. Distribución de muestreo de la razón de dos varianzas

La ley de distribución F

Sean X_1 y X_2 dos variables aleatorias independientes que tienen distribución χ^2 con n_1 y n_2 grados de libertad, respectivamente; entonces la variable aleatoria

$$V = \frac{X_1/n_1}{X_2/n_2}$$

sigue una distribución F (de Snedecor) con (n_1, n_2) grados de libertad, que se la notará como $F(n_1, n_2)$. (Véase la Figura 6.3)

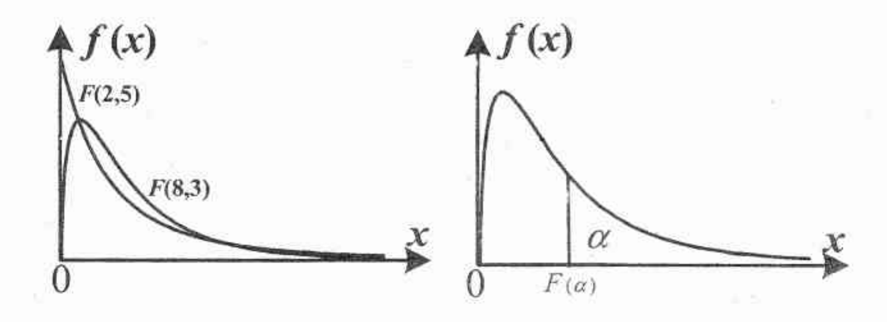


Figura 6.3: Funcion de densidad de la distribución F.

Su función de densidad es

$$f(x) = \begin{cases} \dfrac{\Gamma\left(\dfrac{n_1+n_2}{2}\right) n_1^{n_1/2} n_2^{n_2/2}}{\Gamma\left(\dfrac{n_1}{2}\right)\Gamma\left(\dfrac{n_2}{2}\right)} x^{n_1/2-1}(n_2+n_1x)^{-(n_1+n_2)/2}, & \text{si } x > 0; \\ 0, & \text{si } x \leq 0. \end{cases}$$

La esperanza y la varianza son:

$$\mathbf{E}(V) = \frac{n_2}{n_2-2}, \text{ si } n_2 > 2 \quad \text{y} \quad \operatorname{Var}(V) = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}, \text{ si } n_2 > 4.$$

Nótese que esta ley depende de dos parámetros (n_1, n_2) que corresponden a sus grados de libertad del numerador y del denominador, respectivamente.

Los valores de las probabilidades vienen tabulados. En la tabla 4 del Apéndice se presenta el valor x para el cual la variable aleatoria $V \sim F(n_1, n_2)$ es igual a una probabilidad α: $\Pr(V > x) = \alpha$.

Para la lectura de los valores porcentuales del extremo inferior de la tabla de la ley F se emplea la siguiente relación:

$$F_{(1-\alpha)}(n_1, n_2) = \frac{1}{F_\alpha(n_2, n_1)}.$$

Ejemplos

1. Determinar el valor de x tal que $\Pr(V > x) = 0.05$, donde $V \sim F(6,9)$. En la tabla de la ley F correspondiente a $\alpha = 0.05$, se localiza los valores de $n_1 = 6$ y $n_2 = 9$, para los grados de libertad. Donde se cruzan la columna y la fila correspondientes se lee el valor $x = F_{0.05}(6,9) = 3.37$.

2. Hallar el valor de x tal que $\Pr(V < x) = 0.05$, donde $V \sim F(6,9)$. Aquí, $n_1 = 6$, $n_2 = 9$ y si $\Pr(V < x) = 0.05$, entonces $\Pr(V > x) = 0.95$. Por la relación anterior,

$$F_{0.95}(6,9) = \frac{1}{F_{0.05}(9,6)} = \frac{1}{4.1} = 0.244.$$

La ley de distribución de $\frac{s_1^2}{s_2^2}$

Supongamos que se dispone de dos poblaciones que siguen una ley normal: la población 1 sigue una ley $\mathcal{N}(\mu_1, \sigma_1^2)$ y la población 2 sigue una ley $\mathcal{N}(\mu_2, \sigma_2^2)$. Sean s_1^2 y s_2^2 las varianzas de dos muestras aleatorias independientes de tamaños n_1 y n_2, seleccionadas respectivamente de las poblaciones 1 y 2. Entonces, la variable aleatoria

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

sigue una distribución F con (n_1-1, n_2-1) g.l.

Tengamos presente que si $\sigma_1^2 = \sigma_2^2 = \sigma^2$, entonces $F = \dfrac{s_1^2}{s_2^2} \sim F(n_1-1, n_2-1)$.

Ejemplo. Una marca de automóviles tiene dos plantas que ensamblan el mismo modelo de autos. El rendimiento de estos automóviles debe tener la misma media y desviación estándar. La empresa tiene la política de regularmente comparar los rendimientos de los carros ensamblados, escogiendo

muestras aleatorias en las dos plantas. Se tomaron sendas muestras de tamaño 30 y se controló la desviación estándar del consumo de cada una. Hallar la probabilidad de que la desviación estándar de una muestra sea al menos 1.5 veces mayor que la de la segunda.

Solución: Se tiene que:

$$\Pr\left(\frac{s_1}{s_2} \geq 1.5\right) = \Pr\left(\frac{s_1^2}{s_2^2} \geq 2.25\right).$$

Como $F = \dfrac{s_1^2}{s_2^2} \sim F(29, 29)$; entonces,

$$0.01 < \Pr\left(\frac{s_1^2}{s_2^2} \geq 2.25\right) < 0.025.$$

La probabilidad exacta es $\Pr\left(\dfrac{s_1^2}{s_2^2} \geq 2.25\right) = 0.01632$[1].

6.5. Ejercicios

Distribución de la media

1. Para una prueba de aritmética se sabe, con base en la experiencia, que la puntuación media es 70 puntos con una desviación estándar de 12.5. Si se aplica la prueba a 90 personas seleccionadas al azar, aproxime las siguientes probabilidades:

   a) $\Pr(68.5 < \overline{X} < 71.5)$;
   b) $\Pr(66 < \overline{X} < 74)$;
   c) $\Pr(\overline{X} > 72)$;
   d) $\Pr(\overline{X} < 67.5)$.

2. En una ciudad, el peso de los recién nacidos se distribuye según una ley de media $\mu = 3100\,\text{g}$ y desviación estándar $\sigma = 150\,\text{g}$. Halle los parámetros de la distribución que siguen las medias de las muestras de tamaño 100.

3. Un actuario estableció el siguiente modelo probabilístico sobre los sueldos que reciben los trabajadores en el sector de la agroindustria:

Sueldo	200	300	400	500	600
Pr	0.1	0.2	0.4	0.2	0.1

   Si de este sector, se toman 30 sueldos, al azar,

   a) Halle la esperanza y la varianza de la media muestral;
   b) Calcule la probabilidad de que la media muestral se ubique entre 360 y 430 dólares.

4. Las normas internacionales de calidad indican que los neumáticos deben durar al menos 33 mil km. Un fabricante de neumáticos señala que su producto tiene una duración promedio de 34 mil km y desviación estándar de 4 mil km. En un laboratorio que controla la calidad de fabricación se probaron 36 llantas de esta marca. ¿Cuál es la probabilidad de que, en promedio, los neumáticos probados no cumplan con las normas internacionales?

[1] Este valor se obtuvo mediante el empleo de un programa computacional. Nosotros, por la limitación de las tablas, solo podemos acotar el valor de la probabilidad.

5. El tiempo que los usuarios de una empresa interprovincial de transporte esperan para que su bus salga del terminal es una variable aleatoria con media de 8.2 min y desviación estándar de 5.5 min. Suponga que en un bus se embarcaron 49 pasajeros. Halle la probabilidad de que el tiempo promedio que ellos tuvieron que esperar sea:

   a) menor a 10 min. b) entre 7 y 10 min; c) mayor a 7.5 min.

6. La gente que frecuenta cierto bar tiene una probabilidad de 0.001 de salir y cantar con el grupo que está actuando. En una noche de fin de semana hay 150 personas en el bar. ¿Cuál es la probabilidad de que al menos una persona salga y cante con el grupo? (Suponga que cada persona en el bar toma la decisión independientemente del resto. Halle el verdadero valor y el aproximado)

7. El tiempo de permanencia de los automóviles en un gran parqueadero es una variable aleatoria de media 176 minutos y desviación estándar 40 minutos. Calcule, aproximadamente, la probabilidad de que el tiempo medio de permanencia en el parqueadero de 100 automóviles, elegidos al azar, sea superior a 180 minutos.

8. La estatura de los varones de 18 años de Quito sigue una distribución normal de media 162 cm y desviación estándar 13 cm. Se toma una muestra al azar de 85 de estos chicos encuestados y se calcula el promedio. ¿Cuál es la probabilidad de que este promedio se encuentre entre 159 y 164 cm?

9. El centro de cómputo de su universidad dispone de un servidor para gestionar las páginas web personales de profesores y alumnos. Supongamos que la cantidad de memoria ocupada por una de estas páginas puede considerarse como una variable aleatoria con una media de 1.3 Mb y una desviación estándar de 0.3. Si el servidor va a gestionar un total de 500 páginas, calcule, aproximadamente, la probabilidad de que la cantidad promedio de memoria necesaria supere los 1.32 Mb.

10. Se efectuó un análisis sobre la duración de las máquinas impresoras, de una cierta marca, que tienen las empresas públicas. Se eligió una muestra de 179 máquinas utilizadas en una empresa elegida al azar. La vida media de las impresoras resultó ser de 3.33 años y una desviación estándar de 2.05 años. Con una probabilidad del 99.7 %, ¿en qué intervalo de tiempo puede considerarse que se encuentra la vida media de las impresoras de tal marca?

11. Con una muestra de 160 entrevistas realizadas a mujeres que trabajan, resultó que el gasto promedio mensual en arreglo del cabello fue de 39 dólares y desviación estándar de 5.2 dólares. Con una probabilidad del 99.7 %, ¿entre qué límites variará el gasto medio en arreglo del cabello para las mujeres que trabajan?

12. Un proceso automático llena fundas de chifles cuyo peso medio es de 450 g y una desviación estándar de 3 g. Para controlar el proceso, cada hora se pesan 36 fundas escogidas al azar; si el peso neto está entre 449 g y 451 g se continua con el proceso, en caso contrario se detiene el proceso para recalibrar la máquina.

    a) ¿Cuál es la probabilidad de detener el proceso cuando el peso neto medio realmente es 450 g?;
    b) ¿Cuál es la probabilidad de aceptar que el peso neto promedio es 450 g, cuando realmente es de 448 g?

13. La vida útil de cierta marca de llantas sigue una distribución normal X con media 38 mil km y desviación estándar 3 mil km.

a) Si la utilidad Y (en dólares) que produce cada llanta está dada por la relación $Y = 0.2X + 100$, ¿cuál es la probabilidad de que la utilidad sea mayor que 8900 dólares?;

b) Determine el número de tales llantas que debe adquirir una empresa de transporte para conseguir una utilidad media de al menos 7541 dólares, con una probabilidad de 0.996.

14. En Manabí, el peso de los esposos y de las esposas se distribuye según las leyes $\mathcal{N}(80, 100)$ y $\mathcal{N}(64, 69)$, respectivamente, y son independientes. Si se eligen 25 matrimonios, al azar, de Manabí, calcule la probabilidad de que el promedio de los pesos sea a lo más 137 kg.

Distribución de la proporción

15. Se extrae una muestra aleatoria de 150 elementos de una población binomial con $p = \frac{1}{4}$, ¿cuál es la probabilidad de que la proporción muestral satisfaga $\frac{3}{16} \leq \widehat{p} \leq \frac{5}{16}$?

16. El suceso A tiene una probabilidad de 0.4. Esto significa que esperamos que la frecuencia relativa de A esté cercana a 0.4 en una larga serie de repeticiones del experimento que se está modelando. ¿Cuál es la probabilidad de que en 1000 experimentos, la frecuencia relativa esté entre 0.38 y 0.42 (inclusive)?

17. La FIFA está interesada en conocer si las selecciones nacionales ganan más de la mitad de los partidos que juegan en casa. Suponga que se escogen aleatoriamente los resultados de 80 partidos, efectuados en las más recientes eliminatorias para el Mundial de Fútbol, y se encuentra que 65 % de ellos fueron ganados por el equipo local.

a) ¿Es el 65 % un parámetro o un estadístico? Explique;

b) Asumiendo que no hay ventaja de campo, y por lo tanto que los equipos locales ganan el 50 % de sus juegos, determine la probabilidad de que los equipos locales hubieran ganado el 65 % o más de sus partidos en una muestra de 80 resultados;

c) ¿La información muestral (que el 65 % de los juegos fueron ganados por el equipo de casa) provee fuerte evidencia que los equipos locales ganan más de la mitad sus partidos? Explique.

18. Supongamos que el 80 % de todos los residentes en Guayaquil celebran la fiesta de Navidad (el 25 de diciembre.) Se planea seleccionar una muestra aleatoria de 300 guayaquileños y determinar la proporción de ellos que celebran la Navidad.

a) ¿Es el 80 % un parámetro o un estadístico? ¿Qué símbolo usa para representarlo?;

b) De acuerdo al Teorema del Límite Central, ¿cómo variará la proporción de quienes celebran la Navidad, de muestra a muestra?;

c) Determine la probabilidad que menos de las tres cuartas partes de la muestra celebre la fiesta;

d) ¿La probabilidad calculada en c) sería mayor, menor o igual si el tamaño de la muestra fuera de 800 personas? (Usted no necesita realizar cálculos.) Explique.

19. En un canal de transmisión de datos la probabilidad de que un bits se reciba con un error es 1×10^{-5}. Si en una transmisión se envían 16 millones de bits, ¿cuál es la probabilidad de que no ocurran más de 150 errores?

20. Según las estadísticas de tránsito, se ha establecido que en una noche de viernes, en promedio, 1 de cada 10 conductores está ebrio. Si un fin de semana la policía realiza 400 pruebas de alcolemia, ¿cuál es la probabilidad de que el número de conductores ebrios detectados:

a) menos de 8 %?;

b) más de 12.5 %?;

c) al menos 10 %, pero menos de 13 %?

21. Supongamos que el 40 % de los votantes está a favor de la reelección del actual alcalde.

    a) Si se selecciona una muestra de 600 electores de la ciudad, ¿cuál es la probabilidad de que la proporción muestral de votos a favor del alcalde esté entre el 37 % y el 45 %?;

    b) ¿Cuál debe ser el tamaño muestral para tener una probabilidad del 97 % de que la proporción de votos a favor del alcalde en la muestra no se diferencie de la proporción supuesta en más del 2 %?

22. La mediana de la edad de los habitantes del Ecuador es de 26 años. Si se seleccionan 100 residentes en Ecuador al azar, calcule la probabilidad de que por lo menos el 60 % de ellos tenga menos de 26 años.

23. Se ha estimado que el 43 % de los estudiantes de leyes considera que es muy importante que se imparta un curso de ética en la abogacía. De una población de 800 estudiantes se tomó una muestra de 80. Calcule la probabilidad de que más de la mitad de ellos opinen de ese modo.

24. En la segunda vuelta electoral los resultados dan que el candidato ganador obtuvo el 55 % de los votos. ¿Cuál es la probabilidad de que en una encuesta realizada a 169 personas el resultado no muestre una mayoría a favor del candidato?

25. En una encuesta realizada con una muestra de 3000 personas adultas escogidas al azar, ha resultado que el 35 % toma café al menos una vez al día. Con una probabilidad del 95.5 %, ¿entre qué límites variará esta proporción para la población completa?

26. El tiempo que esperan los peatones para cruzar una vía muy transitada se distribuye en forma exponencial con media de 1 minuto. Si en una hora llegan 95 peatones, calcule la probabilidad de que por lo menos la tercera parte de ellos tenga que esperar más de un minuto.

27. Un ajedrecista experimentado ha ganado el 70 % de las partidas que ha jugado. Si en el próximo mes va a participar en un torneo en el que va a jugar 25 partidas,

    a) calcule la probabilidad aproximada de que gane por lo menos el 80 % de ellas;

    b) calcule la probabilidad binomial exacta de que gane por lo menos 20 partidas;

    c) ¿qué hipótesis son necesarias para que sean válidas las respuestas a) y b)?

Distribución de la varianza

28. Con el empleo de la tabla de la ley χ^2 localice los siguientes valores y represéntelos, aproximadamente:

    a) $\chi^2_{0.95}(9)$; b) $\chi^2_{0.99}(12)$; c) $\chi^2_{0.025}(20)$; d) $\chi^2_{0.05}(16)$.

29. Si $X_1, X_2, \ldots, X_9$ son nueve variables aleatorias independientes y distribuidas según una ley $\mathcal{N}(17, 32)$, calcule la probabilidad de que la varianza muestral sea menor o igual que 56.28.

30. Calcule la probabilidad de que una muestra de tamaño 13 seleccionada de una población normal con varianza 4 tenga una varianza muestral:

    a) menor que 7.01;

    b) entre 1.19 y 2.1.

31. Encuentre la probabilidad de que una muestra aleatoria de 20 observaciones, de una población normal con varianza $\sigma^2 = 5$, tenga una varianza muestral s^2: a) mayor a 8.1; b) entre 2.66 y 9.52.

32. En los últimos 5 años, las calificaciones del examen de aptitud para el ingreso a la universidad, siguen una distribución normal con varianza $\sigma^2 = 8$. ¿Consideraría usted $\sigma^2 = 8$ como un valor válido de la varianza de las notas de los exámenes que se rindieron este año, si una muestra aleatoria de 20 calificaciones arrojó un valor de $s^2 = 16$?

33. En una oficina de selección de aspirantes para optar por una beca se estudia la varianza de las calificaciones para identificar fácilmente a los mejores aspirantes. Para una prueba de matemáticas se supone que las calificaciones se distribuyen normalmente con desviación estándar de 10. Hay 15 aspirantes a optar por una beca. Calcule la probabilidad de que la desviación estándar de las calificaciones de dichos aspirantes sea mayor que 7.

34. En una granja piscícola se mide la variabilidad en el peso de los peces capturados. Las normas internacionales indican que el peso está distribuido según la ley normal con varianza $\sigma^2 = 225\,\mathrm{g}^2$. Se pesan 27 peces y se calcula su varianza s^2.

    a) Estime aproximadamente $\Pr(s^2 > 150)$;
    b) Halle $\Pr(s^2 > 362)$;
    c) Calcule $\mathbf{E}(s^2)$ y $\mathrm{Var}(s^2)$.

35. El sueldo anual de los empleados de una institución se supone que sigue una distribución normal con desviación estándar de 100 dólares. Si en una inspección la oficina de impuestos toma una muestra de los sueldos de 17 empleados, determine un intervalo en el que quedará la varianza, de los sueldos anuales, con una probabilidad de 0.9.

Otras distribuciones de muestreo

36. Con el empleo de la tabla de la ley t localice los siguientes valores y represéntelos, aproximadamente:

    a) $t_{0.1}(9)$; b) $t_{0.01}(12)$; c) $t_{0.025}(20)$; d) $t_{0.05}(16)$.

37. Si X_1, X_2, ..., X_9 son nueve variables aleatorias independientes y distribuidas según una ley $\mathcal{N}(8, 4)$, calcule la probabilidad $\Pr\left(7 \leq \overline{X} \leq 9; 1.09 \leq s^2 \leq 10.045\right)$. ($\overline{X}$ y s^2 son independientes)

38. En la ciudad capital, el precio medio de venta de las casas nuevas es 115 mil dólares. Se toma una muestra aleatoria de 10 casas nuevas, resultando una desviación estándar de 25 mil dólares. ¿Cuál es la probabilidad de que la media muestral de los precios de venta sea:

    a) menor de 104 mil dólares?; b) mayor de 110 500 dólares?

39. Se tomó una muestra de 16 directores de oficinas de una ciudad con el fin de estimar el tiempo medio diario que emplean en desplazarse hasta su trabajo. Si la media de los tiempos es de 87 minutos y la desviación estándar de 20 minutos, calcule la probabilidad de que la media muestral sea menor de 100 minutos.

40. Con el empleo de la tabla de la ley F localice los siguientes valores y represéntelos, aproximadamente:

    a) $F_{0.1}(9, 5)$; b) $F_{0.01}(12, 12)$ c) $F_{0.025}(20, 7)$; d) $F_{0.05}(15, 4)$.

41. Dos muestras aleatorias independientes de tamaños 21 y 9, respectivamente, se toman de una misma población normalmente distribuida. ¿Cuál es la probabilidad de que la varianza de la primera muestra sea al menos el cuádruplo de la varianza de la segunda?

42. Dos muestras aleatorias independientes de tamaños 7 y 13, respectivamente, se toman de una misma población normalmente distribuida. ¿Cuál es la probabilidad de que la varianza de la primera muestra sea mayor igual al triple de la varianza de la segunda muestra?

43. Sean $X_1 \sim \chi^2(9)$, $X_2 \sim \chi^2(20)$ y $Y = \dfrac{X_1/9}{X_2/20}$. Halle los valores de a y b tales que

$$\Pr(a \leq Y \leq b) = 0.925 \quad \text{y} \quad \Pr(Y \leq a) = 0.05.$$

44. Sean X_1 y X_2 una muestra escogida de una población normal estándar.

a) Determine la ley de distribución de $F = \left(\dfrac{X_1 + X_2}{X_1 - X_2}\right)^2$;

b) Calcule la probabilidad $\Pr(F < 16)$.

45. Una muestra aleatoria de tamaño 16 se seleccionó a partir de una población normal de media 75 y desviación estándar 8. Una segunda muestra aleatoria de tamaño 9 se tomó a partir de una población normal de media 70 y desviación estándar 12. Sean $\bar{X}_1$ y $\bar{X}_2$ dos medias muestrales. Halle:

a) la probabilidad de que $\bar{X}_1 - \bar{X}_2$ sea mayor que 4;

b) la probabilidad que $3.5 \leq \bar{X}_1 - \bar{X}_2 \leq 5.5$.

46. Una firma comercializadora afirma que el peso medio (en gramos) μ_1 y μ_2 de dos marcas de atún enlatado, A_1 y A_2, es el mismo. Para verificar la afirmación se escogen dos muestras independientes de tamaños 36 de cada marca. Si la media muestral de A_1 es mayor que la media muestral de A_2, se rechaza que $\mu_1 = \mu_2$; en caso contrario, se acepta que $\mu_1 = \mu_2$. ¿Cuál es la probabilidad de aceptar que $\mu_1 = \mu_2$, cuando realmente $\mu_1 = \mu_2 + 2$? Suponga que las poblacionales son $\sigma_1^2 = 9$ y $\sigma_B^2 = 4$.

47. Para comparar la duración media (en meses) μ_1 y μ_2 de dos marcas de baterías, A y B, se tomaron dos muestras aleatorias independientes de tamaños 32 y 36, respectivamente. Si la duración promedio (muestral) de A es mayor que la de B en más de dos meses, se acepta que $\mu_1 > \mu_2$; caso contrario, se acepta que $\mu_1 = \mu_2$. Calcule la probabilidad de aceptar que $\mu_1 > \mu_2$, cuando realmente $\mu_1 = \mu_2$. Suponga que las varianzas de las duraciones son $\sigma_A^2 = 16$ y $\sigma_B^2 = 9$.

48. El administrador de un edificio quiere decidir la compra de lámparas fluorescentes de marca T o U. Para ayudarle a realizar su decisión, se escogen dos muestras de tamaños 10 y 9 lámparas, respectivamente, resultando las desviaciones estándar de $s_1 = 200$ y $s_2 = 150$. Si la diferencia entre los promedios es mayor que 173 horas, se acepta que $\mu_1 \neq \mu_2$; de lo contrario se acepta que $\mu_1 = \mu_2$. ¿Cuál es la probabilidad de aceptar que $\mu_1 \neq \mu_2$, cuando realmente $\mu_1 = \mu_2$? (Asuma que la vida útil de ambas marcas tiene distribución normal con varianzas iguales.)

49. Para comparar los salarios que pagan a sus empleados dos fábricas de cobijas, San Lucas y Cebra, se escogen dos muestras aleatorias de tamaños 16 y 13, respectivamente, de las dos fábricas. Resultó que las desviaciones estándar fueron $s_L = 120$ dólares y $s_C = 55$ dólares. Si la diferencia entre las medias muestrales no es mayor a 65 dólares, se acepta que $\mu_1 = \mu_2$; caso contrario, se acepta que $\mu_1 \neq \mu_2$. ¿Cuál es la probabilidad de aceptar que $\mu_1 \neq \mu_2$, cuando realmente $\mu_1 = \mu_2$? Suponga que los salarios, en ambas empresas, siguen una distribución normal con varianzas diferentes.

50. Dos programas de televisión tienen como ratings 40 % y 20 %, respectivamente. Se tomó una muestra de 300 hogares que poseen televisor, durante la transmisión del programa A y otra muestra de 100 hogares durante la transmisión de B. ¿Cuál es la probabilidad de que los resultados muestren que el programa A tiene un rating mayor al de B, en un 10 %?

51. Un fabricante afirma que el 30 % de mujeres y el 20 % de hombres prefieren su jabón de tocador. Si se realiza una encuesta a 200 personas de cada sexo, ¿con qué probabilidad la proporción muestral de mujeres menos la proporción muestral de varones están en el intervalo (−19 %, 19 %)?

52. Se escoge una muestra de 600 electores que acaban de votar, entre las 9:00 h y las 15:00 h, para estimar la proporción de votantes a favor de los candidatos H y M. En una encuesta realizada la víspera, se estimó en 30 % y 35 % los porcentajes de apoyo de los dos candidatos, respectivamente. ¿Cuál es la probabilidad de que la proporción muestral de B exceda a la proporción muestral de A en al menos 10 %?

53. La música romántica es preferida por el 30 % de mujeres y el 25 % de hombres. En una encuesta realizada a 300 personas de cada sexo, ¿cuál es la probabilidad de que la proporción muestral de mujeres que prefieren la música romántica, sea mayor a la de los hombres?

Capítulo 7

Estimación de Parámetros

Todos hemos estado haciendo estadística toda la vida,
en el sentido que cada uno ha estado ocupado sacando conclusiones
a partir de observaciones empíricas, casi desde el nacimiento
W. Kruskal

En el Capítulo 1 se expusieron varios métodos que permiten describir un conjunto de datos de manera rápida, general y eficaz; estos métodos son gráficos y su interpretación es fácil, pero tienen el inconveniente que la descripción de los datos no es única y no se prestan para realizar predicciones.

El uso de la información que se obtiene a partir de una muestra para sacar conclusiones sobre la población de la que ella proviene, se denomina *inferencia estadística.*

Los métodos de la inferencia estadística que se describirán en éste y los siguientes capítulos se refieren a la estimación de parámetros, la formulación y verificación de hipótesis sobre estos parámetros y el planteamiento de modelos adecuados para los datos.

7.1. Estimación

La teoría de la estimación de parámetros fue desarrollada en las primeras décadas del siglo XX, como una parte de otra teoría (las pruebas de hipótesis) y sistematizada por J. Neyman en 1934. Actualmente, esta teoría es la base de cualquier estudio estadístico.

Cuando se toma una muestra de una población, el objetivo es tener un indicio de los valores de los parámetros desconocidos de ésta. Tal proceso se denomina *estimación* y a los valores calculados *estimadores.*

Definición (de estimador) Un estimador es una medida estadística que permite conocer o tener una idea del valor de un parámetro desconocido, basándose en la información de la muestra.

Por ejemplo, si disponemos de una población cuya media μ es desconocida, es natural escoger el promedio $\overline{x}$ como estimador de μ.

Observemos que un estimador es una variable aleatoria; mientras que una estimación es un número.

Algunas veces, los estimadores de los parámetros poblacionales se distinguen del verdadero valor mediante el empleo del símbolo $\widehat{\ }$. Por ejemplo,

$p =$ verdadero valor de la proporción poblacional.

$\widehat{p} =$ proporción poblacional estimada, a partir de una muestra.

Las siguientes secciones las dedicaremos a conocer las propiedades de los estimadores de los parámetros poblacionales, a evaluar su validez y a exponer sus aplicaciones.

7.2. Clases de estimadores

Cuando se obtiene una muestra de una población, el objetivo es tomar una decisión en base de los estadísticos calculados a partir de los datos muestrales; luego ellos se resumen en frases como las siguientes:

1. En 2930 de los 10 000 hogares de la ciudad se sintonizaba cierto programa de televisión.
2. El nivel de desocupación en el país es el 14 % de la población en edad de trabajar.
3. La muestra usando el Material A tiene, en promedio, una fortaleza a la tensión 3.2 unidades mayor que aquella empleando el Material B.

Los estimadores anteriores dan una idea concisa de los resultados de la muestra, pero no informan de su precisión. Así, pudiera haber gran diferencia entre tales estimaciones, calculadas a partir de una muestra, y lo que uno podría obtener si dispusiera de una cantidad ilimitada de datos. Por ejemplo, 14 % sería una estimación razonable (o predicción) de la desocupación el próximo mes; pero, ¿cuán «buen» estimador es? Teniendo en cuenta la variación en el mercado laboral, sabemos que es improbable que el próximo mes haya un nivel de desocupación de *exactamente* el 14 %. Sin embargo, podemos esperar que su valor sea «cercano» al 14 %, y ¿qué tan cercano? ¿Podemos esperar que sea dentro del ±0.1 % del estimador?, o ¿dentro del ±1 %?, o ¿dentro del ±10 %?

A partir de la discusión anterior podemos deducir que existen dos tipos de estimadores: uno que da un valor numérico que resume lo observado en la muestra; y otro que, además, expresa la incertidumbre debida a la variabilidad en los (generalmente limitados) datos. A continuación definimos más formalmente estos tipos.

Definición (de estimador puntual) Sea X_1, X_2, ..., X_n una muestra aleatoria seleccionada de una población con distribución de parámetro θ. Se denomina estimador puntual del parámetro θ a cualquier estadístico que proporciona una estimación del verdadero valor de θ.

Al estimador puntual de θ se le nota $\widehat{\theta}_n$.

Por ejemplo, si la media poblacional es $\mu = 6$, obtenemos una muestra y determinamos un promedio $\overline{x} = 5.85$. Ésta es una estimación puntual de μ.

También, mencionamos que la estimación puede realizarse mediante un rango de valores entre los cuales se encontrará el verdadero valor θ con alta probabilidad.

Definición (de estimador por intervalo) Un estimador por intervalo de un parámetro desconocido θ está dado por dos puntos, que pretenden abarcar el valor real del parámetro.

Por ejemplo, si la media poblacional es $\mu = 6$, obtenemos una muestra y determinamos el intervalo $I = (5.4; 6.2)$, este intervalo es una estimación de μ. Posteriormente, indicaremos cómo construir tales intervalos y evaluaremos su exactitud.

7.3. Estimación puntual

No todo estimador realiza una buena estimación del parámetro. Un buen estimador es aquel que está cerca del parámetro estimado; para lo cual debe cumplir ciertas propiedades, que vamos a examinarlas a continuación.

7.3.1. Propiedades de los estimadores puntuales

Lo importante de un estimador es que él pueda emplearse de manera confiable; por ejemplo, que no difiera de manera apreciable del verdadero valor del parámetro poblacional.

Definición(de estimador insesgado) Un estimador $\widehat{\theta}$ es insesgado para estimar θ si

$$\mathbf{E}(\widehat{\theta}) = \theta.$$

De otra manera $\widehat{\theta}$ se llama sesgado.

La Figura 7.1 muestra un estimador insesgado y un estimador sesgado.

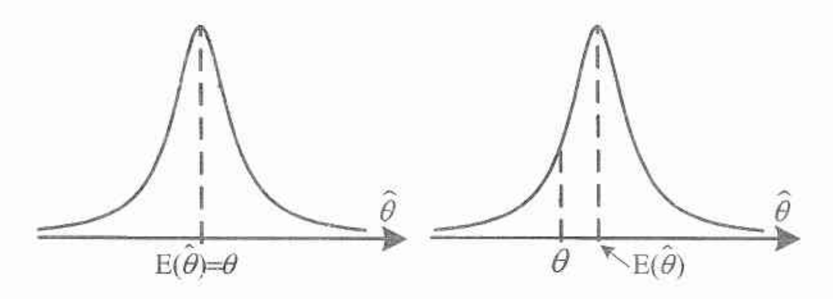


Figura 7.1: Estimadores insesgado y sesgado.

Notemos que la distribución muestral para el estimador sesgado está desplazada hacia la derecha de θ. Este estimador sesgado, probablemente, sobrestima θ.

El sesgo de un estimador se mide como

$$\text{sesgo}(\widehat{\theta}) = \mathbf{E}(\widehat{\theta}) - \theta.$$

Ejemplos

1. La media muestral $\overline{X}$ es un estimador insesgado de μ, ya que $\mathbf{E}\left(\overline{X}\right) = \mu$.

2. El estadístico $\sqrt{\overline{X^2}} = \sqrt{\dfrac{\sum_{i=1}^{n} X_i^2}{n}}$ no es un estimador insesgado de μ, ya que $\mathbf{E}\left(\sqrt{\overline{X^2}}\right) \neq \mu$. ◄

Si disponemos de dos estimadores insesgados de θ, interesa tener un criterio para elegir uno de ellos.

Definición(de estimador eficiente) Sean $\widehat{\theta}_1$ y $\widehat{\theta}_2$ dos estimadores insesgados de θ; $\widehat{\theta}_2$ se dice que es más eficiente que $\widehat{\theta}_1$ si $\text{Var}(\widehat{\theta}_2) < \text{Var}(\widehat{\theta}_1)$. (Ver la Figura 7.2.)

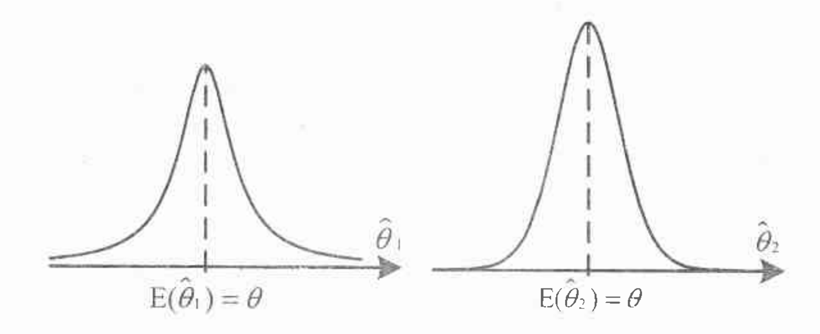


Figura 7.2: Estimadores insesgados con distinta varianza: $\mathrm{Var}(\widehat{\theta}_2) < \mathrm{Var}(\widehat{\theta}_1)$.

Dados dos estimadores insesgados de un mismo parámetro, es preferible escoger el más eficiente.

A veces se presenta el problema de elegir entre dos estimadores con propiedades contrapuestas: uno de ellos es insesgado y el otro es sesgado, pero con menor varianza. En estos casos es necesario definir una medida que nos permita realizar tal comparación.

Definición (de error cuadrático medio) El error cuadrático medio debido a la estimación de θ mediante $\widehat{\theta}$ es $\mathbf{ECM}(\widehat{\theta}) = \mathbf{E}\left[(\theta - \widehat{\theta})^2\right]$.

Por las propiedades de la esperanza se tiene que

$$\mathbf{ECM}(\widehat{\theta}) = \mathrm{Var}(\widehat{\theta}) + \left(\mathrm{sesgo}(\widehat{\theta})\right)^2.$$

Así, para los estimadores insesgados, $\mathbf{ECM}(\widehat{\theta}) = \mathrm{Var}(\widehat{\theta})$.

Dados dos estimadores de un mismo parámetro, es preferible escoger el que tenga el menor **ECM**.

Ejemplo. Sea $X_1, X_2, \ldots, X_5$ una muestra aleatoria de una población para la cual $\mathbf{E}(X_i) = \mu$ y $\mathrm{Var}(X_i) = \sigma^2$, $i = 1, 2, \ldots, 5$. Se tienen los siguientes estimadores de μ:

$$\widehat{\theta}_1 = X_1, \qquad \widehat{\theta}_2 = \frac{1}{3}(X_1 + X_3 + X_5),$$
$$\widehat{\theta}_3 = \frac{1}{2}(X_1 + 2X_5), \qquad \widehat{\theta}_4 = \overline{X} = \frac{1}{5}(X_1 + X_2 + X_3 + X_4 + X_5).$$

a) Según el criterio de la eficiencia, ¿cuál es el mejor?; b) Comparar los estimadores $\widehat{\theta}_2$ y $\widehat{\theta}_3$ mediante el **ECM**.

Solución: Calculemos las esperanzas y las varianzas de cada uno de los estimadores:

$$\mathbf{E}(\widehat{\theta}_1) = \mu, \quad \mathbf{E}(\widehat{\theta}_2) = \mu, \quad \mathbf{E}(\widehat{\theta}_3) = \frac{3}{2}\mu, \quad \mathbf{E}(\widehat{\theta}_4) = \mu.$$

$$\begin{aligned}
\mathrm{Var}(\widehat{\theta}_1) &= \mathrm{Var}(X_1) = \sigma^2, \\
\mathrm{Var}(\widehat{\theta}_2) &= \frac{1}{9}(\mathrm{Var}(X_1) + \mathrm{Var}(X_3) + \mathrm{Var}(X_5)) = \frac{\sigma^2}{3}, \\
\mathrm{Var}(\widehat{\theta}_3) &= \frac{1}{4}(\mathrm{Var}(X_1) + 4\,\mathrm{Var}(X_5)) = \frac{5\sigma^2}{4}, \\
\mathrm{Var}(\widehat{\theta}_4) &= \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{5}.
\end{aligned}$$

a) El estimador $\widehat{\theta}_3$ es sesgado, mientras que $\widehat{\theta}_1$, $\widehat{\theta}_2$ y $\widehat{\theta}_4$ son insesgados; además, $\widehat{\theta}_4 = \overline{X}$ es el mejor estimador de μ, porque tiene la menor varianza (es el más eficiente).

b) Calculemos los sesgos de los dos estimadores:

$$\begin{aligned}\text{sesgo}(\widehat{\theta}_2) &= \mathbf{E}(\widehat{\theta}_2) - \theta = \mu - \mu = 0.\\ \text{sesgo}(\widehat{\theta}_3) &= \mathbf{E}(\widehat{\theta}_3) - \theta = \frac{3}{2}\mu - \mu = \frac{1}{2}\mu.\end{aligned}$$

Entonces, el **ECM** es

$$\begin{aligned}\mathbf{ECM}(\widehat{\theta}_2) &= \text{Var}(\widehat{\theta}_2) + \left(\text{sesgo}(\widehat{\theta}_2)\right)^2 = \frac{\sigma^2}{3} + 0 = \frac{\sigma^2}{3},\\ \mathbf{ECM}(\widehat{\theta}_3) &= \text{Var}(\widehat{\theta}_3) + \left(\text{sesgo}(\widehat{\theta}_3)\right)^2 = \frac{5\sigma^2}{4} + \left(\frac{1}{2}\mu\right)^2 = \frac{5\sigma^2 + \mu^2}{4}\end{aligned}$$

El mejor estimador es $\widehat{\theta}_2$ porque tiene el menor error cuadrático medio. ◀

7.3.2. Estimación de la media poblacional

A continuación demostraremos que el promedio es un estimador insesgado de la media poblacional y calcularemos su varianza.

Supongamos que se obtiene una muestra $X_1, X_2, \ldots, X_n$ de una población que tiene media μ y varianza σ^2. La media muestral se define por $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$.

- La esperanza de $\overline{X}$ es

$$\begin{aligned}\mathbf{E}(\overline{X}) &= \mathbf{E}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n}\mathbf{E}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}(X_i) = \frac{1}{n}(n\mu)\\ &= \mu.\end{aligned}$$

  De manera que $\overline{X}$ es un estimador insesgado de μ.

- La varianza de $\overline{X}$ es

$$\begin{aligned}\text{Var}(\overline{X}) &= \text{Var}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) = \frac{1}{n^2}(n\sigma^2)\\ &= \frac{\sigma^2}{n}.\end{aligned}$$

7.3.3. Estimación de la varianza poblacional

Supongamos que se obtiene una muestra $X_1, X_2, \ldots, X_n$ de una población que tiene media μ y varianza σ^2. El estimador

$$S^{*2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

es un estimador puntual de σ^2, pero tiene el inconveniente de ser sesgado, por lo que se define

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

que es un estimador insesgado para σ^2.

- La esperanza de S^2 es

$$\begin{aligned}\mathbf{E}(S^2) &= \mathbf{E}\left(\frac{\sum\limits_{i=1}^{n}(X_i-\overline{X})^2}{n-1}\right) = \frac{1}{n-1}\mathbf{E}\left(\sum_{i=1}^{n}\left(X_i^2+\overline{X}^2-2\overline{X}X_i\right)\right)\\ &= \frac{1}{n-1}\mathbf{E}\left(\sum_{i=1}^{n}X_i^2 - n\overline{X}^2\right) = \frac{1}{n-1}\left[\sum_{i=1}^{n}\mathbf{E}\left(X_i^2\right) - n\mathbf{E}\left(\overline{X}^2\right)\right]\end{aligned}$$

  Como $\mathbf{E}(X_i^2) = \mu^2+\sigma^2$ y $\mathbf{E}\left(\overline{X}^2\right) = \mu^2 + \frac{\sigma^2}{n}$, se tiene

$$\begin{aligned}\mathbf{E}(S^2) &= \frac{1}{n-1}\left[\sum_{i=1}^{n}(\mu^2+\sigma^2) - n\left(\mu^2+\frac{\sigma^2}{n}\right)\right]\\ &= \frac{1}{n-1}(n\mu^2+n\sigma^2-n\mu^2-\sigma^2) = \frac{1}{n-1}(n-1)\sigma^2\\ &= \sigma^2.\end{aligned}$$

- La varianza de S^2 es $\text{Var}(S^2) = \dfrac{2\sigma^4}{n-1}$ (este resultado no se demostrará ya que su complejidad sale del alcance de esta obra).

7.4. Métodos de estimación puntual

Para determinar la estimación de un parámetro poblacional existen varios métodos, los dos más importantes son el de los momentos y el de máxima verosimilitud.

7.4.1. Método de los momentos

El método de estimación de los momentos fue desarrollado por K. Pearson en 1880. Es un método general de estimación de uno o más parámetros y se basa en la idea de tomar como estimador de la media a la media muestral, como estimador de la varianza a la varianza muestral, y así sucesivamente.

En general, si denotamos $\mu_k = \mathbf{E}(X^k)$, denominado el $k-$ésimo *momento teórico* de la variable X. El $k-$ésimo *momento muestral* es la variable

$$\widehat{\mu}_k = \frac{\sum\limits_{i=1}^{n}X_i^k}{n}, \quad k = 1, 2, 3, \ldots$$

Entonces, igualamos los correspondientes momentos teórico y muestral: $\mu_k = \widehat{\mu}_k$ y resolvemos la **ecuación resultante para $\boldsymbol{\theta}$.**

El inconveniente de este método es que los estimadores obtenidos, muchas veces son sesgados.

Ejemplo. Disponemos de una muestra X_1, X_2, ..., X_n, proveniente de una población con distribución $\mathcal{U}[-\theta, \theta]$, donde $\theta > 0$ es desconocido. Hallar la estimación de θ.

Solución: Tenemos que $\mu = \mathbf{E}(X) = 0$ y $\mu_2 = \mathbf{E}\left(X^2\right) = \mathrm{Var}(X) + (\mathbf{E}(X))^2 = \dfrac{\theta^2}{3}$. Por otro lado, $\widehat{\mu}_2 = \dfrac{\sum\limits_{i=1}^{n} X_i^2}{n}$.

Si igualamos μ_2 y $\widehat{\mu}_2$, obtenemos $\dfrac{\theta^2}{3} = \dfrac{\sum\limits_{i=1}^{n} X_i^2}{n}$.

Entonces, el estimador de θ es $\widehat{\theta} = \sqrt{\dfrac{3}{n}\sum\limits_{i=1}^{n} X_i^2}$.

7.4.2. Método de máxima verosimilitud

El método de estimación por máxima verosimilitud fue desarrollado por R. A. Fisher en 1920. La ventaja de este método es que utiliza la información sobre la ley de distribución de la que proviene la muestra.

Sea X_1, X_2, ..., X_n una muestra proveniente de una distribución con parámetro θ y ley $f(x;\theta)$. El procedimiento a seguir es el siguiente:

1. Determinar la *función de verosimilitud* de la muestra en sus valores respectivos: Esta función del parámetro θ está dada por
$$L(\theta) = f(X_1;\theta) \times f(X_2;\theta) \times \cdots \times f(X_n;\theta) = \prod_{i=1}^{n} f(X_i;\theta).$$
2. Encontrar la *función de log-verosimilitud* tomando el logaritmo de la función de verosimilitud.
$$l(\theta) = \log(L(\theta)) = \sum_{i=1}^{n} \log f(X_i;\theta).$$
3. Hallar el valor de θ que maximiza la log-verosimilitud. En este caso, es el valor $\widehat{\theta}$, que es solución de la ecuación
$$\frac{dl}{d\theta} = 0.$$

Observación. Si la distribución de probabilidad contiene k parámetros, θ_1, θ_2, ..., θ_k, la estimación de máxima verosimilitud de cada uno será la solución de las ecuaciones respectivas:
$$\frac{\partial l}{\partial \theta_1} = 0, \quad \frac{\partial l}{\partial \theta_2} = 0, \quad \ldots \quad \frac{\partial l}{\partial \theta_k} = 0.$$

Ejemplo. Sea X_1, X_2, ..., X_n una muestra proveniente de una población con distribución $\mathcal{N}(\mu, \sigma^2)$. Hallar los estimadores de μ y de σ^2.

Solución: La función de densidad de la ley es $f(X;\mu,\sigma^2) = \dfrac{1}{\sqrt{2\pi}\sigma}\exp\left(-\dfrac{(X-\mu)^2}{2\sigma^2}\right)$.

De manera que la función de verosimilitud es

$$\begin{aligned} L(\mu, \sigma^2) &= (f(X;\mu,\sigma^2))^n = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(X_i-\mu)^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\sum_{i=1}^{n} \frac{(X_i-\mu)^2}{2\sigma^2}\right) \end{aligned}$$

La función de log-verosimilitud es

$$l(\mu, \sigma^2) = \log(L(\mu, \sigma^2)) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2.$$

Derivando l esta función respecto a μ, e igualando a cero da

$$\frac{\partial l}{\partial \mu} = -0 + \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu) = 0,$$

de donde $\widehat{\mu} = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X}$.

Por otro lado,

$$\frac{\partial l}{\partial(\sigma^2)} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i-\mu)^2 = 0$$

cuya solución es $\widehat{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n} = S^{*2}$.

7.5. Ejercicios

1. «El principio subyacente en todas las técnicas de inferencia estadística es que uno usa estadísticos muestrales para aprender algo (es decir, para inferir algo) acerca de los parámetros poblacionales». Si usted entendió lo que quiere decir esta afirmación, escriba un párrafo en el que describa una situación en la que se pueda emplear un estadístico muestral para inferir algo sobre un parámetro poblacional. En su ejemplo, identifique claramente la muestra, la población, el estadístico y el parámetro. Sea tan específico como sea posible y no use ejemplos dados en el libro.

2. Se toma una muestra de tamaño 4 de una población de media μ y varianza σ^2. Se propone los siguientes estimadores de la media:

$$\widehat{\theta}_1 = \frac{X_1 + X_2 + 3X_4}{5}, \qquad \widehat{\theta}_2 = \frac{X_1 + X_2 + 2X_3}{4},$$
$$\widehat{\theta}_3 = \frac{X_1 + X_2 + X_3 + X_4}{4}, \qquad \widehat{\theta}_4 = \frac{X_1 + X_2 + X_3 + X_4 - 3}{3}.$$

Indique su orden de preferencia (del mejor al peor) y explique los motivos de su clasificación.

3. Dos muestras aleatorias independientes se extraen de una población con media μ y varianza σ^2. Los tamaños muestrales son n_1 y $n_2 = \frac{n_1}{2}$ y las medias muestrales son $\overline{X}_1$ y $\overline{X}_2$, respectivamente. Para estimar a μ se proponen tres estimadores,

$$T_1 = \overline{X}_1, \quad T_2 = \overline{X}_2, \quad T_3 = \frac{\overline{X}_1 + \overline{X}_2}{2}.$$

   a) Diga si los estimadores son insesgados;
   b) Encuentre su varianza;
   c) ¿Cuál de los tres estimadores es mejor?

4. Si se dispone de una muestra X_1, X_2, X_3 de observaciones que siguen una ley exponencial $\mathcal{E}(1/\theta)$. Considere los siguientes estimadores

$$\widehat{\theta}_1 = X_1, \quad \widehat{\theta}_2 = \frac{X_1 + X_3}{2}, \quad \widehat{\theta}_3 = \frac{X_1 + 2X_2}{2}, \quad \widehat{\theta}_4 = \overline{X}.$$

   a) ¿Cuáles estimadores son insesgados para θ?;
   b) Entre los estimadores insesgados de θ, ¿cuál estimador escogería usted y por qué?

5. Si se dispone de una muestra X_1, X_2, X_3, X_4, X_5 de observaciones que siguen una ley de Poisson $\mathcal{P}(\lambda)$. Considere los siguientes estimadores:

$$\widehat{\theta}_1 = \frac{X_1 + X_2 + X_4 + X_5}{4}, \qquad \widehat{\theta}_2 = \frac{X_1 + 2X_2 + X_3 + 2X_4 + X_5}{7},$$
$$\widehat{\theta}_3 = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}, \qquad \widehat{\theta}_4 = X_1^2 + X_2^2.$$

   a) ¿Cuáles estimadores son insesgados para λ?;
   b) Escoja el mejor estimador insesgado de λ.

6. A partir de una población que tiene media μ y varianza σ^2 se tomaron tres muestras de tamaño $n_1 = 7$, $n_2 = 14$ y $n_3 = 9$. Sean s_1^2, s_2^2 y s_3^2 las varianzas muestrales calculadas a partir de las muestras. Compruebe que

$$s^2 = \frac{7s_1^2 + 14s_2^2 + 9s_3^2}{30}$$

   es un estimador insesgado de σ^2.

7. El número de clientes que ingresan a una librería en una hora es una variable aleatoria X que sigue una distribución de Poisson con media λ. Se dispone de una muestra aleatoria $X_1, \ldots, X_n$ del número de clientes por hora.

   a) Determine un estimador insesgado de λ;
   b) Las ventas en una hora se da por $V = 4X + X^2$. Demuestre que $\mathbf{E}(V) = 5\lambda + \lambda^2$;
   c) Determine un estimador insesgado de $\mathbf{E}(V)$ empleando toda la muestra $X_1, \ldots, X_n$.

8. Si X_1 y X_2 son dos variables aleatorias independientes tales que $\mathbf{E}(X_1) = \mathbf{E}(X_2) = \mu$ y $\text{Var}(X_1) = \text{Var}(X_2) = \sigma^2$, determine si el estimador.

$$Z = \frac{(X_1 - X_2)^2}{2}$$

   es un estimador insesgado para σ^2.

9. Suponga que $\widehat{\theta}_1$ y $\widehat{\theta}_2$ son estimadores insesgados de θ y que $\text{Var}(\widehat{\theta}_1) = \sigma_1^2$ y $\text{Var}(\widehat{\theta}_2) = \sigma_2^2$. Se forma un nuevo estimador de θ mediante $\widehat{\theta}_3 = \alpha\widehat{\theta}_1 + (1-\alpha)\widehat{\theta}_2$, $(0 \leq \alpha \leq 1)$. Si $\widehat{\theta}_1$ y $\widehat{\theta}_2$ son independientes,

   a) verifique que $\widehat{\theta}_3$ es insesgado;
   b) ¿Cómo seleccionará α para reducir $\text{Var}(\widehat{\theta}_3)$ al mínimo?

10. Suponga que $\overline{X}_1$ y $\overline{X}_2$ son dos medias muestrales calculadas a partir de dos muestras de tamaño n_1 y n_2, respectivamente, obtenidas de una población normal de media μ y varianza σ^2.

    a) Se define un estimador de μ: $\overline{X}_3 = \alpha\overline{X}_1 + (1-\alpha)\overline{X}_2$, $(0 \leq \alpha \leq 1)$. Pruebe que es insesgado;
    b) Halle el valor de α que minimiza la varianza.

11. Considere una sucesión de variables aleatorias X_1, X_2, ..., X_n independientes e idénticamente distribuidas que siguen una ley uniforme en el intervalo $[\theta - 1; \theta + 1]$, donde θ es un parámetro desconocido. Denotamos por $\overline{X}_n$ el promedio de estas variables.

    a) Calcule $\mathbf{E}(\overline{X}_n)$ y $\text{Var}(\overline{X}_n)$;
    b) Construya un estimador T_n del parámetro θ que sea consistente e insesgado y calcule su error cuadrático medio.

12. Sea X_1, ..., X_n una muestra aleatoria de una población con la siguiente distribución discreta:

$$\Pr(X = 1) = \frac{2(1-\theta)}{2-\theta}, \quad \Pr(X = 2) = \frac{\theta}{2-\theta}, \quad \theta \in (0, 1).$$

    a) Halle el estimador de θ por el método de los momentos;
    b) Halle el estimador de θ por máxima verosimilitud;
    c) Demuestre que los dos estimadores encontrados coinciden.

13. Sea X_1, ..., X_{25} una muestra aleatoria de tamaño 25 una población binomial de parámetro $p \in (0, 1)$.

$$\Pr(X = x) = \mathbf{C}_2^x p^x (1-p)^{2-x}.$$

    Determine el estimador de máxima verosimilitud si el valor 0 ocurre 5 veces, el valor 1 ocurre 11 veces y el valor 2 ocurre 9 veces.

14. Sea X_1, ..., X_n una muestra aleatoria de

$$f(x) = \begin{cases} e^{-x+\theta}, & \text{si } x > \theta, \quad (\theta \in \mathbf{R}); \\ 0, & \text{caso contrario.} \end{cases}$$

    a) Calcule $\mathbf{E}(X)$ y obtenga un estimador de θ mediante el método de los momentos;
    b) Determine si el estimador es insesgado;
    c) Obtenga el estimador de θ por el método de máxima verosimilitud;
    d) ¿Cuál de los dos estimadores es más eficiente?
    e) Calcule el **ECM** de $\widehat{\theta}_2$.

15. El tiempo, en horas, que dura un elemento electrónico es una variable aleatoria T que tiene distribución exponencial de parámetro λ. Para estimar λ se prueban 30 elementos y se encuentra que 18 fallan antes de las 600 horas de uso.

a) Mediante el método de máxima verosimilitud, estimar la proporción de todos los elementos que fallan antes de las 600 horas;

b) Utilice el resultado de a) para obtener un estimador de λ.

16. A lo largo de un año, la hembra de tigrillo puede tener una o dos crías, o no tener ninguna. Según un estudio realizado por un grupo de zoólogos, la proporción de hembras sin crías es $\frac{p}{3}$, la de las con una cría es $\frac{2p}{3}$, mientras que la proporción de las hembras con dos crías es $1-p$, donde el parámetro p toma un valor entre 0 y 1.

a) Halle el número medio esperado de crías por hembra a lo largo de un año;

b) Al realizar un estudio de 200 hembras durante un año, el equipo de científicos mencionado encontró 55 hembras que no han tenido crías, 106 que han tenido una cría y 39 que han tenido dos. Estime el parámetro p por el método de los momentos.

17. El control de una partida de rodillos se realiza clasificando las piezas en pequeñas, normales y grandes. Las proporciones teóricas se suponen $p_1 = 0.05$; $p_2 = 0.90$; $p_3 = 0.05$. Pero se sospecha que ha aumentado la dispersión y, por tanto, las piezas siguen las proporciones: $p_1 = 0.05 + a$; $p_2 = 0.90 - 3a$; $p_3 = 0.05 + 2a$. Se analizan 5000 piezas obteniéndose $n_1 = 1278$; $n_2 = 2928$; $n_3 = 794$ de cada clase. Obtenga una estimación de máxima verosimilitud de a.

7.6. Estimación por intervalo

Recordemos lo que se indicó en una sección anterior: los intervalos estadísticos expresan la incertidumbre debida a la variabilidad de los datos muestrales. Además, para que ellos tengan validez en su aplicación práctica, tienen que cumplir con algunas hipótesis básicas, como que la muestra se debe obtener de manera aleatoria, independiente y debe estar idénticamente distribuida, lo que no siempre se logra, incrementando la mencionada incertidumbre.

Así, por ejemplo, con base en una muestra de hogares en los que se está viendo televisión, podemos construir un intervalo que contenga, con un grado específico de confiabilidad, la media o la desviación estándar del tiempo que la población consultada ve televisión .

Antes de analizar los diferentes casos, examinemos algunas ideas preliminares.

Definición (de intervalo de confianza) Un intervalo de confianza es un rango de valores, calculado a partir de los datos muestrales, el cual probablemente incluye el verdadero valor de un parámetro desconocido.

A cada intervalo de confianza se le asocia una probabilidad $(1-\alpha)$ de que contenga el verdadero valor del parámetro θ. A tal probabilidad se le denomina *nivel de confianza* y a los extremos del intervalo, *límite inferior* y *límite superior* de confianza: $(LIC; LSC)$. Esto se resume en

$$\Pr(LIC \leq \theta \leq LSC) = 1 - \alpha.$$

Un intervalo que cumple estas condiciones se denomina *intervalo de confianza* de nivel $(1-\alpha) \times 100\,\%$.

Para tener resultados fiables, el nivel de confianza debe ser alto, lo más cercano a uno; generalmente se toma 0.9, 0.95 o 0.99.

El ancho de un intervalo de confianza nos da la idea de cuanta incertidumbre existe, alrededor del parámetro desconocido. Un intervalo muy ancho puede indicar que deberíamos recolectar más datos antes de decir algo definitivo sobre el parámetro.

7.7. Estimación de la media cuando la varianza es conocida

Suponga que se desea estimar la media μ de una población cuya varianza σ^2 es conocida y que para tal efecto se dispone de una muestra de n mediciones: x_1, x_2, ..., x_n.

Un intervalo de confianza para la media poblacional μ, a un nivel del $100(1-\alpha)\,\%$, está dado por

$$\left(\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}};\ \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right). \tag{7.1}$$

Donde:

- n es el tamaño de la muestra.
- σ la desviación estándar de la población.
- $z_{\alpha/2}$ el valor z que corresponde al área $\frac{\alpha}{2}$ en el extremo superior de la distribución normal estándar; es decir, $1 - \Phi\left(z_{\alpha/2}\right) = \frac{\alpha}{2}$ (Figura 7.3).

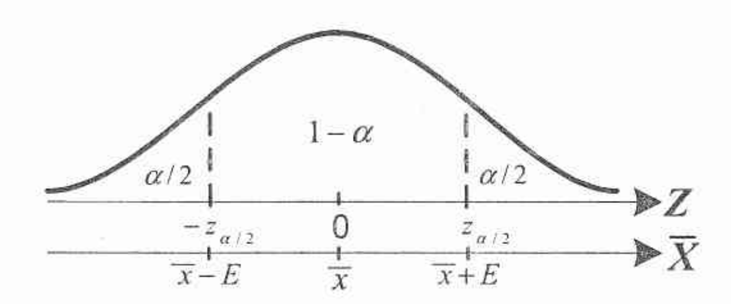


Figura 7.3: Intervalo de estimación para μ.

Aquí se aplica el Teorema del Límite Central y es aconsejable tener un tamaño muestral $n \geq 25$.

En la siguiente tabla se presenta los intervalos de confianza más comunmente usados:

Nivel de confianza	α	$z_{\alpha/2}$	*LIC*	*LSC*
0.90	0.10	1.645	$\overline{x} - 1.645\,\sigma/\sqrt{n}$	$\overline{x} + 1.645\,\sigma/\sqrt{n}$
0.95	0.05	1.96	$\overline{x} - 1.96\,\sigma/\sqrt{n}$	$\overline{x} + 1.96\,\sigma/\sqrt{n}$
0.99	0.01	2.58	$\overline{x} - 2.58\,\sigma/\sqrt{n}$	$\overline{x} + 2.58\,\sigma/\sqrt{n}$

Observaciones

1. Si el tamaño de la muestra es suficientemente grande ($n \geq 25$) y se desconoce la varianza, se puede utilizar el intervalo 7.1, reemplazando σ por su estimador s, sin pérdida de exactitud.

2. Puesto que para un valor de $z_{\alpha/2} = 3$ se tiene un nivel de confianza del 99.7 %, en las aplicaciones prácticas se supone que un intervalo de confianza al 99.7 % contiene el valor de la media, *con toda seguridad.*

Ejemplos

1. Determinar un intervalo de confianza de nivel 95 % para la media poblacional μ si $n = 36$, $\overline{x} = 15$ y $\sigma^2 = 3.24$.

   Solución: Sabemos que $\sigma^2 = 3.24$, o sea $\sigma = 1.8$.

   La confiabilidad es $1-\alpha = 0.95$; entonces, $\alpha = 0.05$ y $\alpha/2 = 0.025$. De la tabla de la distribución normal $z_{\alpha/2} = z_{0.025} = 1.96$.

   De manera que el intervalo es

$$\begin{aligned}\left(\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) &= \left(15 - 1.96\frac{1.8}{\sqrt{36}}; 15 + 1.96\frac{1.8}{\sqrt{36}}\right)\\ &= (14.412; 15.588).\end{aligned}$$

2. En un estudio a trabajadores informales se seleccionó una muestra aleatoria de 50 individuos en una semana dada. Para cada uno, se determinó el número de horas semanales de trabajo efectivo y se obtuvo un promedio de 46 horas. Si se asume que la desviación estándar es 3.64 horas, estimar la media del número de horas trabajadas por todos los informales, con un nivel de confianza del 98 %.

   Solución: Se tiene que $n = 50$, $\overline{x} = 46$ y $\sigma = 3.64$.

   Además, $1 - \alpha = 0.98$, $\alpha = 0.02$, $\alpha/2 = 0.01$ y $z_{0.01} = 2.33$.

   El intervalo resulta:

$$\begin{aligned}\left(\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) &= \left(46 - 2.33\frac{3.64}{\sqrt{50}}; 46 + 2.33\frac{3.64}{\sqrt{50}}\right)\\ &= (46 - 1.2; 46 + 1.2)\\ &= (44.8; 47.2).\end{aligned}$$

   Ello significa que, con una probabilidad del 98 %, la media del número de horas trabajadas aquella semana se encuentra entre 44.8 y 47.2.

3. Según los consumidores, las empresas pasteurizadoras de leche no entregan la cantidad exacta de producto. Para verificar esta denuncia, se tomó una muestra de 45 fundas, cuyo contenido teórico era de 1 litro de leche. Se encontró un promedio de $972\,\text{cm}^3$ y una desviación estándar de $51\,\text{cm}^3$. Sobre la base de un intervalo de confianza al 99.7 %, ¿se puede decir que la denuncia de los consumidores tiene fundamento?

   Solución: Tenemos que $n = 45$, $\overline{x} = 972$ y $s = 51$.

   Si $1 - \alpha = 0.997$, entonces $z_{\alpha/2} = 3$.

   Aunque el valor de la varianza es desconocida, podemos emplear el intervalo de confianza, ya que el tamaño de la muestra es suficientemente alto. De manera que el intervalo es

$$\begin{aligned}\left(\overline{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}; \overline{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right) &= \left(972 - 3\frac{51}{\sqrt{45}}; 972 + 3\frac{51}{\sqrt{45}}\right)\\ &= (949.2; 994.8).\end{aligned}$$

   Como el nivel de confiabilidad es del 99.7 %, podemos decir que *con toda seguridad*, la denuncia de los consumidores es verdadera.

Determinación del tamaño de la muestra

Cuando se estima la media, mediante el promedio, el *error de estimación* viene dado por la distancia entre su valor verdadero y la estimación: $|\overline{x} - \mu|$. Sus valores varían entre 0 y $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

Podemos plantearnos el problema de encontrar el tamaño de la muestra de manera que el error de estimación no sea mayor que E.

El intervalo de confianza para la media poblacional tiene la forma $(\overline{x} - E; \overline{x} + E)$, donde $E = |\overline{x} - \mu|$ es el error en la estimación de la media para el nivel de confiabilidad dado. Si el intervalo tiene la forma $\left(\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$, entonces $E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

Si de esta igualdad despejamos n, obtenemos

$$n = \left(\frac{z_{\alpha/2}\,\sigma}{E}\right)^2,$$

que es el tamaño de la muestra, necesario para tener un error de estimación E a un nivel de confianza $1 - \alpha$.

Ejemplo. Se desea conocer la distancia media que corren semanalmente un grupo de atletas de fondo. Por estudios anteriores se conoce que la desviación estándar de esas distancias es de 3 km. ¿A cuántos atletas habrá que muestrear si la estimación debe quedar a menos de 0.5 km de la media verdadera, con un nivel de confianza del 95 %?

Solución: El intervalo de confianza es de la forma $(\overline{x} - 0.5; \overline{x} + 0.5)$, entonces $E = 0.5$ y $z_{\alpha/2} = 1.96$.

Aplicando la fórmula,

$$n = \left(\frac{z_{\alpha/2}\,\sigma}{E}\right)^2 = \left(\frac{1.96 \times 3}{0.5}\right)^2 = 138.3.$$

Consecuentemente, se debe tener un tamaño de muestra mínimo de 139 atletas. Si la muestra es demasiado alta, se deberá aumentar el error admisible E o disminuir el nivel de confianza $1 - \alpha$.

Se sugiere que el lector ajuste los parámetros, para tener una muestra aproximada de 50 atletas, y los interprete.

7.8. Ejercicios

1. Determine los intervalos de confianza al 95 % para la media poblacional desconocida si

   a) $n = 45$, $\overline{x} = 5$, $\sigma = 6.8$;
   b) $n = 100$, $\overline{x} = 37$, $s^2 = 25$·
   c) $n = 130$, $\overline{x} = 18.5$, $\sigma^2 = 4.2$;
   d) $n = 169$, $\overline{x} = -22$, $\sigma^2 = 14$.

2. Determine la longitud del intervalo de confianza para la media poblacional si:

   a) $1 - \alpha = 0.96$, $n = 25$, $\sigma = 0.86$;
   b) $1 - \alpha = 0.99$, $n = 152$, $\sigma = 128$;
   c) $1 - \alpha = 0.955$, $n = 71$, $\sigma = 3.5$;
   d) $\alpha = 0.03$, $n = 120$, $\sigma^2 = 49$.

3. Obtenga un intervalo de confianza de $100(1 - \alpha)$ % para la media poblacional si

a) $\alpha = 0.01$, $n = 26$, $\overline{x} = 120$, $\sigma = 4$;

b) $\alpha = 0.05$, $n = 65$, $\overline{x} = 222$, $\sigma^2 = 51$;

c) $\alpha = 0.1$, $n = 90$, $\overline{x} = 66$, $s^2 = 2.5$;

d) $\alpha = 0.04$, $n = 55$, $\overline{x} = -37$, $\sigma^2 = 18$.

4. Determine un intervalo en el que se pueda decir que se encuentra el valor de la media con casi toda seguridad si

a) $n = 36$, $\overline{x} = 100$, $\sigma = 4.2$;

b) $n = 44$, $\overline{x} = 53$, $\sigma^2 = 71$;

c) $n = 81$, $\overline{x} = 86$, $s^2 = 22.5$;

d) $n = 121$, $\overline{x} = -37$, $\sigma^2 = 84.1$.

5. La cantidad mínima requerida para que un anestésico surta efecto en una intervención quirúrgica fue, por término medio, de 50 mg, con una desviación estándar de 10.2 mg, en una muestra de 60 pacientes. Obtenga un intervalo de confianza para la media al 97 %, suponiendo que la muestra se extrajo mediante muestreo aleatorio simple sobre una población normal. Interprete el resultado.

6. En cierto barrio se seleccionó, al azar, una muestra de 100 personas cuyo promedio de ingresos mensuales es $\overline{x} = 460$ dólares y una desviación estándar de $\sigma = 200$ dólares.

a) Si se toma un nivel de confianza del 97 %, ¿cuál es el intervalo de confianza para la media de los ingresos mensuales de toda la población?;

b) Si se toma un nivel de confianza del 99 %, ¿cuál es el tamaño muestral necesario para estimar la media de ingresos mensuales con un error menor a 30 dólares?

7. Se tomó una muestra aleatoria de 88 individuos a los que se midió el nivel de glucosa en la sangre, obteniendo una media muestral de $110\,\text{mg}/\text{cm}^3$. Se sabe que la desviación estándar de la población es $20\,\text{mg}/\text{cm}^3$.

a) Obtenga un intervalo de confianza para el nivel de glucosa en sangre de la población, al 90 % de confianza;

b) ¿Qué error máximo se comete con la estimación anterior?

8. La media de edad de los alumnos que se presentan a las pruebas de acceso a la universidad es de 18.1 años y la desviación estándar 0.6 años. De los alumnos se elige, al azar, una muestra de 120.

a) ¿Cuál es la probabilidad de que la media de edad de la muestra esté comprendida entre 17.95 y 18.25 años?;

b) ¿Qué tamaño debe tener una muestra de dicha población para que su media esté comprendida entre 17.9 y 18.3 años, con una confianza del 99.5 %?

9. Una fábrica produce varillas de hierro con una desviación estándar de 25 cm. La empresa recibe un pedido de varillas que indica que la longitud promedio debe tener una desviación máxima de 10 cm de la longitud requerida. ¿Cuántas varillas se tendrán que producir para cumplir con la especificación, con casi toda seguridad?

10. Se realizaron 169 mediciones del voltaje de la red de alumbrado público y se registró un promedio de 108 voltios y desviación estándar de 5 voltios.

a) ¿Cuáles son los límites de confianza, a un nivel del 98 %, para el voltaje medio de red de alumbrado público?;

b) ¿A qué nivel de confianza puede decirse que la estimación de la media incluye el valor 109 voltios?

11. En una región geográfica, la estatura de los individuos varones (en cm) sigue una $\mathcal{N}(\mu; 7.5^2)$.

    a) Halle el intervalo de confianza al nivel 92 % para estimar μ, a partir de una muestra aleatoria de tamaño 36, cuya estatura promedio es 167.2 cm;

    b) Para la misma población, determine el tamaño mínimo de la muestra para estimar μ con un error inferior a ± 2 cm con un nivel de confianza del 94 %.

12. A partir de la información suministrada por una muestra aleatoria de 90 locales comerciales de la ciudad se determinó el intervalo de confianza al 96 %: $I = (42, 58)$, para el gasto medio mensual por familia (en dólares) en electricidad. Determine (justificando las respuestas):

    a) La estimación puntual que daríamos para el gasto mensual por familia en electricidad en esa ciudad;

    b) ¿Qué número de familias tendríamos que seleccionar al azar como mínimo para garantizarnos, con una confianza del 96 %, una estimación de dicho gasto medio con un error máximo no superior a 3 dólares?

13. La vida activa (en días) de cierto fármaco sigue una distribución $\mathcal{N}(1200; 40^2)$. Se desea enviar un lote de medicamentos de modo que la vida media del lote no sea inferior a 1190 días, con probabilidad 0.95. Halle el tamaño mínimo del lote.

14. Se desea conocer el nivel de consumo medio con una determinada tarjeta de crédito con un error máximo de 15 dólares y un nivel de confianza de 0.97. ¿Cuál debe ser el tamaño mínimo de la muestra que se debe tomar, si se ha estimado una desviación estándar de 45 dólares?

15. Se sabe que el contenido de fructosa de una variedad de manzana sigue una distribución normal cuya varianza es conocida teniendo un valor de 0.25. Se desea estimar el valor de la media poblacional mediante el valor de la media de una muestra, admitiendo un error máximo de 0.18, con una confianza del 95.5 %. ¿Cuál es el tamaño de la muestra?

16. Suponga que se midió la longitud del pie derecho a 41 estudiantes de su universidad. El promedio de todas las mediciones fue de 28.4 cm y la desviación estándar fue 5.1 cm.

    a) Encuentre un intervalo de confianza al 96 % para la longitud media del pie derecho de todos los estudiantes de su universidad;

    b) ¿Esperaría usted que alrededor del 96 % de todos los estudiantes tengan longitudes de pies en este intervalo? Explique;

    c) Si usted hubiera encontrado un intervalo de confianza al 90 %, ¿cómo habría diferido del intervalo antes obtenido?;

    d) Si la muestra hubiera constado de 141 estudiantes (los restantes datos se mantienen iguales), ¿cómo habría variado el intervalo de confianza?;

    e) Si la desviación estándar hubiera sido de 3.7 cm (sin variar los demás datos), ¿cómo se hubiera visto afectado el intervalo de confianza?;

    f) Si la media muestral hubiera resultado ser de 25.4 cm (sin variar los demás datos), ¿qué habría sucedido con el intervalo de confianza?

7.9. Estimación de la media cuando la varianza es desconocida

Supóngase que se desea estimar la media μ para una población cuya varianza σ^2 es desconocida y que se dispone de una muestra de n mediciones que siguen una ley normal: $x_1, x_2, \ldots, x_n$.

Un intervalo de confianza para la media poblacional μ, a un nivel del $100(1-\alpha)\,\%$, está dado por

$$\left(\overline{x} - t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}};\ \overline{x} + t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right).$$

Donde:

- s es la desviación estándar muestral.
- $t_{\alpha/2}(n-1)$ es el valor, de la distribución t de Student a $(n-1)$ grados de libertad, para el cual el área en el extremo superior es igual a $\frac{\alpha}{2}$.

Ejemplos

1. La cotización diaria de una moneda frente al dólar sigue una distribución normal de media y varianza desconocidas. Se eligieron 9 días al azar, la cotización fue:

$$65.3 \quad 66.2 \quad 65.8 \quad 66.0 \quad 66.1 \quad 64.5 \quad 65.2 \quad 67.1 \quad 64.2.$$

a) Determine un intervalo de confianza, al 99 %, para la cotización media de la citada moneda;

b) ¿Con qué confiabilidad se estima la media en un intervalo cuya longitud es 1.116?

Solución: De acuerdo a los datos, $n = 9$, $\overline{x} = 65.6$ y $s = 0.9$.
Además, $1 - \alpha = 0.99$; de modo que $\frac{\alpha}{2} = 0.005$ y $t_{0.005}(8) = 3.355$.

a) El intervalo es

$$\begin{aligned}\left(\overline{x} - t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}};\overline{x} + t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right) &= \left(65.6 - 3.355\frac{0.9}{\sqrt{9}};65.6 + 3.355\frac{0.9}{\sqrt{9}}\right)\\ &= (64.59; 66.61).\end{aligned}$$

b) Si la longitud del intervalo es 1.116, entonces el error máximo es $E = \frac{1.116}{2} = 0.558$ y como $E = t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}$; entonces, sustituyendo valores:

$$0.558 = t_{\alpha/2}(8)\frac{0.9}{\sqrt{9}}.$$

De donde, $t_{\alpha/2}(8) = \frac{0.558 \times \sqrt{9}}{0.9} = 1.86$.
Si examinamos en la tabla de la ley t a 8 g.l., vemos que $t_{0.05}(8) = 1.86$. De manera que $\frac{\alpha}{2} = 0.05$; por lo tanto, $\alpha = 0.1$ y $1 - \alpha = 0.9$. El nivel de confianza es del 90 %.

2. El tiempo que un aparato de televisión permanece encendido sigue una ley normal. Por medio de un audímetro se registró este dato en 10 hogares durante una semana y se calculó el promedio diario (en minutos). Los resultados son los siguientes:

248	177	110	200	135	185	220	155	180	166.

a) Determinar un intervalo de confianza para el tiempo promedio diario que los hogares miran televisión, a un nivel del 95 %;

b) Manteniendo la desviación estándar constante, hallar el tamaño mínimo de la muestra para que el error de la estimación sea menor a 20 min.

Solución:

a) El promedio y la desviación estándar son iguales a $\overline{x} = 177.6$ y $s = 39.92$, respectivamente; y si $1 - \alpha = 0.95$, $t_{0.025}(9) = 2.262$.
El intervalo requerido queda como

$$\left(177.6 - 2.262\frac{39.92}{\sqrt{10}}; 177.6 - 2.262\frac{39.92}{\sqrt{10}}\right) = (149.06; 206.14).$$

b) Aquí no se puede aplicar directamente la fórmula del cálculo del tamaño muestral, ya que los valores de la ley t varían según el número de grados de libertad, pero el principio es el mismo.
Sea $E = t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}$ y como éste debe ser menor que 20, se tiene $t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}} < 20$.
Entonces,

$$\begin{aligned} t_{\alpha/2}(n-1) &< \frac{20}{39.92}\sqrt{n} \\ t_{\alpha/2}(n-1) &< 0.5\sqrt{n} \end{aligned}$$

Formemos una tabla en la que se tenga los dos miembros de la desigualdad. El tamaño mínimo de la muestra es el correspondiente al menor valor de n para el cual se cumple la desigualdad.

n	$0.5\sqrt{n}$	$t_{\alpha/2}(n-1)$
10	1.581	2.262
16	2.000	2.131
17	2.061	2.120
18	2.121	2.110

Se observa que para el valor de $n = 18$ se empieza a cumplir la desigualdad; es decir, el tamaño mínimo de la muestra es 18 hogares.

7.10. Ejercicios

1. Encuentre un intervalo de confianza de nivel $(1-\alpha)$ para la media si

a) $n = 12$, $\overline{x} = 3.18$, $s^2 = 0.21$, $\alpha = 0.05$;
b) $n = 26$, $\overline{x} = 46$, $s^2 = 4.8$, $\alpha = 0.1$;
c) $n = 18$, $\overline{x} = 2.5$, $s^2 = 31.5$, $\alpha = 0.01$;
d) $n = 9$, $\overline{x} = -132$, $s = 58$, $\alpha = 0.05$.

2. Se desea estudiar el gasto semanal de fotocopias, en dólares, de los estudiantes universitarios. Se eligió una muestra de 9 estudiantes, elegidos al azar, resultando los gastos:

2.0 2.5 1.9 0.7 1.1 2.1 3.0 0.8 0.7.

Se supone que la variable aleatoria objeto de estudio sigue una distribución normal de media desconocida. Determine el intervalo de confianza del 95 % para la media del gasto semanal en fotocopias por estudiante.

3. En los entrenamientos de un nadador se mide el tiempo que emplea en recorrer los 100 metros libres durante seis días consecutivos. Se han obtenido los siguientes tiempos promedio de cada uno de los días de la semana en que entrenó:

Tiempo	51.2	49.8	51.6	50.4	50.8	50.5

   a) Si se considera que los tiempos se distribuyen normalmente, encuentre un intervalo de confianza, de nivel 95 %, para el tiempo promedio invertido;

   b) ¿Puede esperarse que el nadador rebaje su tiempo de 50 segundos? ¿Por qué?

4. Las tensiones de rotura (en Kp) de 5 cables de acero fueron

$$660 \quad 460 \quad 540 \quad 580 \quad 550.$$

   Suponiendo normalidad para las tensiones, estime la tensión media de rotura mediante un intervalo de confianza al nivel 99 %.

5. Se desea estimar el tiempo medio de ejecución de un programa. Para ello se ejecutó dicho programa 8 veces utilizando conjuntos de datos elegidos aleatoriamente, obteniéndose que la media muestral y la desviación estándar muestral son, respectivamente, 230 ms y 14 ms. Obtenga un intervalo de confianza al 90 % para la media. (Suponga normalidad.)

6. En una entidad de crédito para la microempresa se desea conocer la deuda media de los clientes que tienen préstamos. Los siguientes datos corresponden a la deuda, en dólares, de 16 clientes que se seleccionaron de manera aleatoria.

306	508	299	537	404	210	347	529
514	255	343	476	536	521	659	438

   Si el monto de cada préstamo es una variable aleatoria normal, obtenga los intervalos de confianza del 90 %, 95 % y 99 %, para la deuda media.

7. Al fumigar los productos agrícolas existe el riesgo de que se coloque demasiada cantidad de agroquímicos, con el consiguiente riesgo para el consumidor. De un lote se extrajo una muestra de 10 tomates y se midió la concentración (en μg/ kg de producto) de fosfatos que ellos contenían, resultando:

$$24.9 \quad 23.5 \quad 26.1 \quad 28.2 \quad 26.4 \quad 23.8 \quad 25.1 \quad 25.3 \quad 27.7 \quad 26.6.$$

   a) Determine un intervalo de confianza del 95 % para la concentración;

   b) Si la concentración máxima permitida es de 24 μg/ kg, ¿puede esperarse que el lote sea aceptado para el consumo humano?

8. En una fábrica de conservas se mide las impurezas en un lote destinado a la exportación. En una muestra de 12 frascos de mermelada se obtuvo los siguientes porcentajes de impurezas:

$$2.3 \quad 1.9 \quad 2.1 \quad 2.8 \quad 2.3 \quad 3.6 \quad 1.4 \quad 1.8 \quad 2.1 \quad 3.2 \quad 2.0 \quad 1.9.$$

   a) Asumiendo que las mediciones están normalmente distribuidas, encuentre el intervalo de confianza al 95 %;

   b) Si el porcentaje máximo de impurezas permitido para la exportación es de 1.5 %, ¿se aceptará el lote para ser exportado?

9. El tiempo de vida en cautiverio de 8 especímenes de un tipo de insecto fue de 228 horas, con una desviación estándar de 7 horas.

a) Estime el tiempo de vida promedio para un nivel del 99 %;

b) Responda la pregunta suponiendo que las muestras fueron de 12 y 100 especímenes.

10. Una máquina produce artículos cuya dimensión se controla mediante la toma de una muestra aleatoria. Un día se obtuvieron las siguientes mediciones:

3.7 3.4 3.5 3.5 3.4 3.7 3.9 3.7 3.5 3.7 3.8.

a) Halle el intervalo de confianza para la media a un nivel de confiabilidad de nivel 90 %;

b) Si la media teórica del proceso es de 3.5 y sin cambiar los otros datos. ¿Cuál es el tamaño mínimo de la muestra, para que a un nivel del 90 %, se pueda asegurar que la máquina produce artículos con dimensión igual a la media teórica?

11. Los siguientes son los tiempos, medidos en días laborables, que demoraron 16 trámites de jubilación en el IESS, elegidos al azar:

Días							
159	280	362	222	264	224	101	212
170	485	250	379	179	168	260	149

a) Bajo la suposición de que los tiempos se distribuyen normalmente, determine un intervalo de confianza al 99 % para el tiempo medio de un trámite de jubilación;

b) El director del Instituto ha indicado que los trámites no se demoran más de 180 días. ¿Es razonable suponer que el tiempo medio verdadero es mayor que lo indicado por el director?

12. La siguiente lista contiene la longitud (número de letras en las palabras) para una muestra de 26 palabras del libro Rayuela de Julio Cortázar:

10	2	3	7	2	9	4	8	2	1	7	5	2
5	4	5	12	2	9	4	2	5	2	3	4	7

a) Calcule el promedio y la desviación estándar de la longitud de las palabras;

b) Construya un intervalo de confianza al 99 % para la longitud media de las palabras en Rayuela;

c) Si el tamaño muestral fuera mayor (y el promedio y la desviación estándar fueran los mismos), ¿cómo cambiaría el intervalo de confianza?;

d) Si el promedio fuera mayor (manteniéndose el tamaño muestral y la desviación estándar), ¿cómo cambiaría el intervalo de confianza?;

e) Un intervalo de confianza al 95 % es (3.655; 6.037). ¿Qué proporción de las 26 palabras de la muestra están dentro del intervalo? ¿Su respuesta será siempre cercana al 95 %? Explique.

7.11. Estimación de la varianza (distribución normal)

Supóngase que se desea estimar la varianza poblacional y que para el efecto se dispone de una muestra de n mediciones que siguen una ley normal: $x_1, x_2, \ldots, x_n$.

Un intervalo de confianza para la varianza poblacional σ^2, a un nivel del $100(1-\alpha)\,\%$, está dado por

$$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}; \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}\right).$$

Donde:

- s^2 es la varianza muestral.
- $\chi^2_{\alpha/2}(n-1)$ es el valor, de la distribución χ^2 a $(n-1)$ grados de libertad, para el cual el área en el extremo superior es igual a $\frac{\alpha}{2}$.
- $\chi^2_{1-\alpha/2}(n-1)$ es el valor, de la distribución χ^2 a $(n-1)$ grados de libertad, para el cual el área en el extremo inferior es igual a $\frac{\alpha}{2}$.

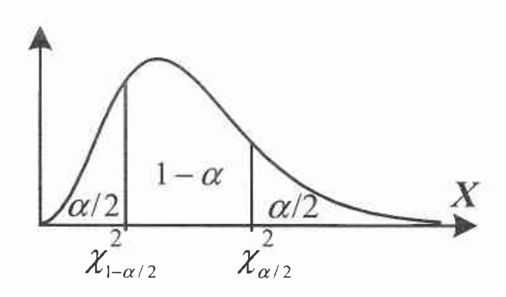


Figura 7.4: Localización de los valores de la ley χ^2 en el intervalo de confianza para σ^2.

Ejemplos

1. Hallar un intervalo de confianza para la varianza poblacional, al 90 %, para una muestra de tamaño $n = 10$, si $s^2 = 196$.

   Solución: Si $1 - \alpha = 0.9$, entonces $\alpha = 0.1$, $\frac{\alpha}{2} = 0.05$ y $1 - \frac{\alpha}{2} = 0.95$. Los valores de $\chi^2_{0.05}$ y $\chi^2_{0.95}$ correspondientes a $n - 1 = 9$ g.l. son

$$\begin{aligned} \chi^2_{\alpha/2}(n-1) &= \chi^2_{0.05}(9) = 16.919, \\ \chi^2_{1-\alpha/2}(n-1) &= \chi^2_{0.95}(9) = 3.325. \end{aligned}$$

   El intervalo de confianza es

$$\begin{aligned} \left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}; \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}\right) &= \left(\frac{9 \times 196}{16.919}; \frac{9 \times 196}{3.325}\right) \\ &= (104.26; 530.53). \end{aligned}$$

2. Un hombre de negocios está interesado en invertir en un instrumento que piensa le puede dar altos rendimientos. No obstante, sabe que, en general, a mayor rendimiento se tiene mayor riesgo. Al considerar instrumentos similares se observaron los siguientes rendimientos porcentuales, que suponemos siguen una ley normal:

$$8.7 \quad 15.5 \quad 21.0 \quad 18.9 \quad 17.3 \quad 22.1 \quad 18.6.$$

   El inversionista considera que si se tiene un riesgo mayor a 10 (desviación estándar), no le conviene invertir. Construir el intervalo de confianza del 95 % y decir si este hombre hará la inversión o no.

   Solución: A partir de los datos, se obtiene que $s^2 = 19.67$, $n = 7$ y $1 - \alpha = 0.95$.

   En la tabla de la ley χ^2 a 6 g.l. se lee:

$$\begin{aligned} \chi^2_{\alpha/2}(n-1) &= \chi^2_{0.025}(6) = 14.45, \\ \chi^2_{1-\alpha/2}(n-1) &= \chi^2_{0.975}(6) = 1.24. \end{aligned}$$

El intervalo de confianza es:

$$\begin{aligned}\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}; \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}\right) &= \left(\frac{6\times 19.67}{14.45}; \frac{6\times 19.67}{1.24}\right)\\ &= (8.17; 95.18).\end{aligned}$$

Es decir, $\sigma^2 \in (8.17;\ 95.18)$; por lo que $\sigma \in \left(\sqrt{8.17}; \sqrt{95.18}\right) = (2.86; 9.76)$.

Como el límite superior del intervalo de confianza es menor a 10, posiblemente si le conviene invertir.

7.12. Ejercicios

1. Halle un intervalo de confianza al 90 % si $s^2 = 225$ para los siguientes tamaños de muestra:

   a) $n = 4$; b) $n = 8$; c) $n= 13$; d) $n = 20$.

2. La varianza de la presión sanguínea es importante porque ella permite conocer la respuesta cardiaca ante el esfuerzo. Se llevaron a cabo 12 mediciones de la presión de un individuo, que dieron los siguientes resultados:

   116 105 121 119 110 105 108 102 107 102 104 116.

   Determine los intervalos de confiabilidad del 90 % y del 95 % para σ^2.

3. En los manuales de fábrica de un aparato para determinar el nivel de alcohol en la sangre, se indica que las mediciones tienen una desviación estándar de 5 unidades. Para probar esta afirmación, en la oficina de normas se realizaron mediciones del contenido de alcohol en la sangre de 10 voluntarios, con los siguientes resultados:

   83 75 92 79 60 85 92 77 76.

   Basándose en un intervalo de confianza de nivel 95 %, determine si la afirmación del fabricante es correcta.

4. Se anlizó una marca de margarina dietética para determinar el nivel de acidos grasos polisaturados (en porcentaje). Una muestra de seis paquetes proveyó la siguiente información.

   16.8 17.2 17.4 16.9 16.5 17.1.

   Asumiendo que los datos siguen una ley normal, halle:

   a) el intervalo de confianza para el contenido medio de grasa, con un nivel del 99 %;

   b) el intervalo de confianza para la varianza del contenido de grasa, con un nivel del 95 %.

5. En una muestra aleatoria de 15 cuentas bancarias que realizaron depósitos la última semana se encontró que la desviación estándar era de 73.6 dólares. Se supone que los depósitos siguen una ley normal. Estime la varianza y la desviación estándar de los depósitos mediante un intervalo de confianza al 95 %.

6. Un hombre de negocios está interesado en invertir en bonos de un país latinoamericano, que piensa le pueden dar altos rendimientos. No obstante, sabe que, en general, a mayor rendimiento se tiene mayor riesgo. Al considerar papeles similares se observaron los siguientes rendimientos (%):

$$9.7 \quad 21.5 \quad 17.0 \quad 16.9 \quad 13.4 \quad 20.1 \quad 24.6 \quad 15.8.$$

El inversionista considera que si se tiene un riesgo mayor a 10 (desviación estándar), no le conviene invertir. Construya el intervalo de confianza del 95 % y diga si esta persona hará la inversión o no, justificando su respuesta.

7. El contenido en nicotina de los cigarrillos de una marca determinada sigue una distribución $\mathcal{N}(\mu; \sigma^2)$. Se tomó una muestra de 5 cigarrillos, obteniéndose en esta muestra un contenido medio de 21.2 mg y varianza muestral 4.2025. Obtenga:

a) el intervalo de confianza para μ, con un nivel del 90 %;

b) el intervalo de confianza para σ^2, con un nivel del 95 %.

8. Se desea probar un nuevo método de embalaje de mercaderías, para lo cual se registra el tiempo (en segundos) que un mismo trabajador emplea en realizar la tarea:

Procedimiento tradicional	Procedimiento alternativo
31	36
36	32
34	30
27	26
40	35
43	41
34	28
30	33
33	

Si el tiempo empleado es similar en los dos casos, entonces se decidirá emplear aquel método que presente la menor variación. Mediante un intervalo de confianza al 95 %, ¿cuál de los dos métodos escogería, el tradicional o el alternativo?

9. Se mide el tiempo (en segundos) de duración de un proceso químico realizado 20 veces en condiciones similares, obteniéndose los siguientes resultados:

93	90	97	90	93	91	96	94	91	88
93	95	91	89	92	87	88	90	86	91

Suponiendo que la duración sigue una distribución normal, hallar los intervalos de confianza al 90 % para la media y la varianza.

10. En el embalaje de fruta para la exportación es importante conocer la variabilidad del calibre de la fruta (que es el diámetro máximo de la fruta). Una fruta con un calibre bajo se cotiza a bajo precio y una con calibre alto da problemas en el embalaje. Las siguientes mediciones corresponden al calibre promedio (en cm) de los melones contenidos en 13 cajas:

20.8	19.3	19.5	22.4	21.3	19.6	20.5
21.2	21.4	19.0	19.8	22.2	20.6	

a) Determine los intervalos de confianza, al 90 %, para la media y la varianza del calibre;

b) Si el calibre promedio es menor que 20 cm o mayor que 21.5 cm, o si la varianza es mayor o igual a 3, se recomienda el cambio de la variedad de melón. ¿Cree usted que será necesario hacer tal cambio? ¿Por qué?

11. Una empresa de venta de cosméticos está interesada en introducir una nueva línea de artículos, para ello se examina la ganancia (en dólares) que le dejarían cada uno de los productos:

23 38 15 7 12 10 10 11 18 13 9 10

a) Calcule un intervalo de confianza para la ganancia media, a un nivel del 95 %;

b) Calcule un intervalo de confianza para la varianza de la ganancia, a un nivel del 90 %;

c) La empresa introducirá la nueva línea de productos si la ganancia media es mayor que 20.5 o si la varianza es menor que 40. En base a los intervalos antes encontrados, ¿se introducirán al mercado los nuevos productos? Explique por qué.

7.13. Estimación de la proporción (distribución binomial)

Históricamente, este intervalo de confianza es, con seguridad, el primero propuesto para un parámetro; en los escritos de Laplace (*Théorie Analytique de Probabilités*, 1812, pp. 283) ya se lo encuentra analizado.

Suponga que se dispone de una muestra $x_1, x_2, \ldots, x_n$, de n observaciones que siguen una ley de Bernoulli, cuyo parámetro p (la proporción poblacional) deseamos estimar.

Un intervalo de confianza aproximado para la proporción p, a un nivel de $100(1-\alpha)\,\%$ viene dado por

$$\left(\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}; \widehat{p} + z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right).$$

Donde:

- $\widehat{p} = \dfrac{y}{n}$, siendo $y = \sum_{i=1}^{n} x_i$ el número de éxitos en las n pruebas.
- $z_{\alpha/2}$ el valor z que corresponde al área $\dfrac{\alpha}{2}$ en el extremo superior de la distribución normal estándar.

Determinación del tamaño de la muestra

Si notamos como $E = |\widehat{p} - p|$ al error en la estimación de la proporción, para el nivel de confiabilidad dado; entonces, $E = z_{\alpha/2}\sqrt{\dfrac{\widehat{p}(1-\widehat{p})}{n}}$.

El tamaño de la muestra necesaria para tener un error E, a un nivel de confianza $(1-\alpha)$ es

$$n = \frac{(z_{\alpha/2})^2\,\widehat{p}(1-\widehat{p})}{E^2}.$$

Observación. Si no se conoce de antemano una estimación de p –como sucede cuando se realiza una investigación por primera vez–, se toma $\widehat{p} = 0.5$, porque este valor permite obtener el tamaño máximo de muestra.

Ejemplos

1. Con el objeto de estimar la proporción de televidentes que han visto el anuncio de un producto, se entrevistó a 400 telespectadores y resultó que 344 de ellos lo habían visto.

   a) Encuentre un intervalo de confianza de 91 % para la proporción de todos los espectadores que han visto la publicidad del producto;

   b) Obtenga el tamaño de muestra indispensable para que el intervalo del inciso a) tenga una longitud máxima de 6 % con la misma confianza.

   Solución: Por el enunciado se tiene: $y = 344$, $n = 400$, $1 - \alpha = 0.91$ y $z_{0.045} = 1.695$.

   La proporción de espectadores en la muestra es $\widehat{p} = \dfrac{y}{n} = \dfrac{344}{400} = 0.86$.

   a) El intervalo de confianza es

   $$\left(\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}; \widehat{p} + z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right) =$$

   $$\left(0.86 - 1.695\sqrt{\frac{(0.86)(0.14)}{400}}; 0.86 + 1.695\sqrt{\frac{(0.86)(0.14)}{400}}\right) = (0.8306; 0.8894).$$

   b) Si la longitud del intervalo es 6 %, quiere decir que $2E = 0.06$; es decir, $E = 0.03$.
   El tamaño de la muestra es

   $$n = \frac{(z_{\alpha/2})^2\,\widehat{p}(1-\widehat{p})}{E^2} = \frac{(1.695)^2(0.86)(0.14)}{(0.03)^2} = 384.35.$$

   Habrá que consultar a 385 televidentes.

2. En una encuesta piloto, previa para la realización de la encuesta definitiva, se encontró que el 63 % de la población cree que el principal problema del país es la corrupción. La ficha técnica de la encuesta definitiva indica que el sondeo tendrá un 97 % de confiabilidad y el error estimado del 4 %. A cuántos ciudadanos se deberá consultar si:

   a) se usa el valor estimado de p mediante la encuesta piloto;

   b) no se emplea una estimación previa de p.

   Solución:

   a) Según los datos: $1 - \alpha = 0.97$, $E = 0.04$ y $\widehat{p} = 0.63$. Entonces,

   $$\begin{aligned} n &= \frac{(z_{\alpha/2})^2\widehat{p}(1-\widehat{p})}{E^2} = \frac{(z_{0.015})^2(0.63)(0.37)}{(0.04)^2} \\ &= \frac{(2.17)^2(0.63)(0.37)}{0.0016} = 686. \end{aligned}$$

   La encuesta deberá ser realizada a un mínimo de 686 personas.

   b) Como no se tiene información previa sobre p, se toma $\widehat{p} = 0.5$ y el cálculo del tamaño queda:

   $$\begin{aligned} n &= \frac{(z_{\alpha/2})^2\widehat{p}(1-\widehat{p})}{E^2} = \frac{(2.17)^2(0.5)(0.5)}{(0.04)^2} \\ &= 735.77. \end{aligned}$$

   Sin un conocimiento previo de la proporción, se deberá muestrear al menos a 736 personas.

En el Cuadro 7.1 se encuentra un resumen de los intervalos de confianza de una muestra analizados en este capítulo.

	Hipótesis	Parámetro	Intervalo
Distribución general	σ conocido	media μ	$\overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
Distribución normal	σ desconocido	media μ	$\overline{x} \pm t_{\alpha/2;(n-1)}\frac{s}{\sqrt{n}}$
Distribución normal		varianza σ^2	$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2;(n-1)}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2;(n-1)}}\right)$
Distribución binomial		proporción p	$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$

Cuadro 7.1: Intervalos de confianza comunmente empleados

7.14. Ejercicios

1. Determine los intervalos de confianza para la proporción, de una muestra de tamaño 200 en la cual se han obtenido 150 éxitos, según los siguientes niveles:

   a) $\alpha = 5\,\%$; b) $\alpha = 12\,\%$; c) $\alpha = 1\,\%$; d) $\alpha = 20\,\%$.

2. En esta pregunta no realice cálculos, responda mediante una frase que explique su razonamiento.

   a) Tres investigadores, Luis, Juan y Édgar, seleccionaron independientemente muestras de una misma población. Los tamaños de las muestras fueron de 4000 para Luis, 1000 para Juan y 250 para Édgar. Cada uno construyó un intervalo de confianza para p a partir de sus datos. Los errores E para los tres intervalos son 0.015, 0.031 y 0.062. Asigne cada uno de los errores a los investigadores;

   b) Dos investigadoras, Diana y Oliva, seleccionaron dos muestras de tamaño 1000 a partir de diferentes poblaciones y construyeron los intervalos de confianza al 95 % para p. El error para el intervalo de Diana fue de 0.030 y el error para el de Oliva fue de 0.025. Dado que las proporciones muestrales fueron de $\widehat{p} = 0.2$ y $\widehat{p} = 0.4$, asigne a cada una de las investigadoras su proporción;

   c) Un investigador seleccionó aleatoriamente 100 sujetos de una población, observó 50 éxitos y calculó tres intervalos de confianza. Los niveles de confiabilidad son 90 %, 95 % y 99 %. y los intervalos son A: (0.402; 0.598), B: (0.371; 0.629) y C: (0.418; 0.582). Asigne a cada intervalo su nivel de confianza;

   d) Dos investigadores, un hombre y una mujer, trabajan juntos para estudiar una muestra aleatoria de individuos de una población. Ellos encuentran que la proporción muestral es de $\widehat{p} = 0.6$. Cuando ellos construyen el intervalo de confianza basado en la proporción, el varón obtiene (0.532; 0.668), mientras que la mujer obtiene (0.552; 0.688). Indique quién se equivocó.

3. Una muestra realizada a los clientes de un supermercado dio que 120 de 300 clientes usan regularmente tarjeta de crédito o cheques para sus compras. Encuentre un intervalo de confianza al 98 % para el porcentaje de personas que usan efectivo en sus compras.

4. Un partido político que concurre a las elecciones municipales en la ciudad quiere encargar una encuesta para estimar su porcentaje de votación mediante un intervalo de la forma $P \pm 1.5\,\%$. cuyo nivel de confianza sea 95 %. ¿Qué tamaño muestral debe utilizarse en la encuesta para alcanzar aproximadamente este objetivo, sabiendo que en una muestra piloto el porcentaje de votación estimado fue del 15 %?

5. La efectividad de un medicamento contra el dolor de cabeza se examina determinando si éste elimina o no el síntoma. Se administró el medicamento a 225 pacientes voluntarios, de los cuales en 170 causó el efecto deseado. El medicamento se acepta para su uso general si tiene una efectividad en al menos el 80 % de los casos.

   a) Basándose en un intervalo de confianza del 98 %, ¿puede recomendarse el uso del medicamento?;
   b) Sin variar la proporción estimada, ¿qué tan grande deberá ser una muestra si se desea tener una confianza del 96 % de que el error máximo de estimación es 0.05?

6. Según un estudio sobre los niños que padecen dolor de pecho, realizado por Selbst, Ruddy y Clark (*Clinical Pediatrics*, 1990), se encontró que de 137 niños que tenían dolor de pecho, 100 daban radiografías de tórax normales.

   a) Obtenga un intervalo de confianza del 95 % para la proporción p de niños con dolor de pecho que dieron radiografías normales;
   b) Halle el mínimo tamaño muestral para que el error cometido en la estimación de p sea inferior a 0.07, al nivel de 95 %.

7. En una población, nadie es indiferente respecto a la iniciativa propuesta por el alcalde de construir un nuevo parque en el norte de la ciudad. Cada habitante adulto o bien está a favor, o bien en contra de la iniciativa. Se desea conocer el porcentaje (P) de las personas que están en contra. Entre 250 habitantes adultos elegidos al azar, 75 afirmaron que estaban en contra (y los 175 restantes a favor).

   a) Halle el intervalo de confianza al 93 % para P;
   b) ¿Cuál es el número mínimo de encuestados necesario para que el error cometido en la estimación sea, como mucho, de 0.05?

8. En una línea de control de calidad en un día se examinan 250 piezas de un lote, de las cuales 25 tienen algún tipo de defecto.

   a) Construya un intervalo de confianza al 93 % para la proporción de piezas buenas y de piezas defectuosas;
   b) Un lote se acepta si la proporción de piezas defectuosas es menor o igual al 5 %. ¿Será aceptado el lote en cuestión?;
   c) ¿Qué tamaño deberá tener una muestra si se desea tener una confianza del 97 % de que la estimación estará dentro del 1 % del porcentaje real?

9. Para la introducción al mercado de una nueva variedad de semilla de maíz la empresa productora estima que deben germinar al menos el 73 % del total de semillas sembradas. En una prueba de laboratorio se sembraron 745 semillas, de las cuales germinaron 518. Con base en un intervalo de confianza de nivel igual al 97 %, ¿podrá la empresa introducir al mercado la nueva variedad?

10. En un sondeo sobre la preferencia deportiva de la población masculina ecuatoriana, realizada a 1000 personas, se determinó que el 72 % de los encuestados gustaba ver regularmente partidos de fútbol por televisión. Con una seguridad del 95 %, ¿se puede decir que los resultados son iguales, con un margen de más o menos 3 puntos porcentuales, a los que se habrían obtenido si se hubiera consultado la opinión de la población masculina completa?

11. Una noticia de prensa dice que, de 1200 personas encuestadas sobre la conveniencia de hacer reformas a la ley de tránsito, 756 se muestran a favor y 444 en contra, y concluye afirmando que el 63 % de la población se muestra a favor, con un margen de error de $\pm 3\,\%$. ¿Cuál es el nivel de confianza de esta afirmación?

12. En una entrevista realizada a 130 mujeres casadas, 113 de ellas indicaron que habían sido víctimas de algún tipo de agresión por parte de su cónyuge.

    a) Asumiendo que estas mujeres forman una muestra aleatoria, calcule un intervalo de confianza de nivel 95 % para la proporción de las mujeres casadas que han sido agredidas;
    b) Si se hubiera consultado a 520 mujeres, ¿cree usted que el intervalo hubiera sido más ancho, más estrecho o de igual ancho? Explique y no realice cálculos;
    c) ¿El intervalo hubiera sido más ancho, más estrecho o de igual ancho si 73 de las 130 mujeres hubieran respondido afirmativamente? Explique;
    d) Realice una interpretación del intervalo.

13. Se desea estimar la proporción de estudiantes universitarios a favor de sustituir el actual himno nacional por otra canción.

    a) Para estimar esta proporción con una precisión de 0.10 a un nivel de confianza del 92 %, ¿a cuántos estudiantes se necesitará preguntar? (Para determinar el tamaño de muestra necesario, fije su propia proporción muestral, identificándola claramente.)

       Para responder las siguientes preguntas, usted no necesita realizar cálculos. Explique sus respuestas.
    b) Si se deseara estimar la proporción con una precisión del 0.05, a un nivel del 92 %, ¿es necesario muestrear a más o menos estudiantes que en a)?;
    c) Si se deseara estimar la proporción con una precisión del 0.02, a un nivel del 95 %, ¿es necesario muestrear a más o menos estudiantes que en a)?

7.15. Intervalos de confianza para dos muestras

Los intervalos de confianza que se presentan a continuación involucran a dos muestras y, en general, son utilizados para comparar entre los valores que tienen los parámetros de dos poblaciones o para determinar si las muestras provienen de la misma población.

7.15.1. Intervalo de confianza para la diferencia entre dos medias

Suponga que disponemos de dos muestras independientes de tamaños n_1 y n_2, seleccionadas de dos poblaciones con medias μ_1 y μ_2. Nuestro interés es encontrar un intervalo de confianza para la diferencia de las medias y para ello consideraremos 3 casos:

Caso 1. Las varianzas poblacionales son conocidas

Supondremos que las varianzas de las poblaciones 1 y 2 son conocidas: σ_1^2 y σ_2^2.

Un intervalo de confianza para la diferencia de medias poblacionales $\mu_1 - \mu_2$, a un nivel del $100(1-\alpha)\,\%$, está dado por

$$\left((\overline{x}_1 - \overline{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}};\ (\overline{x}_1 - \overline{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right).$$

Caso 2. Las varianzas poblacionales son iguales: $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Un intervalo de confianza para la diferencia de medias poblacionales $\mu_1 - \mu_2$, a un nivel del $100(1-\alpha)\,\%$, está dado por

$$\left((\overline{x}_1 - \overline{x}_2) - t_{\alpha/2}(n_1+n_2-2)\sqrt{\frac{s^2}{n_1}+\frac{s^2}{n_2}};\ (\overline{x}_1 - \overline{x}_2) + t_{\alpha/2}(n_1+n_2-2)\sqrt{\frac{s^2}{n_1}+\frac{s^2}{n_2}}\right).$$

Donde

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}.$$

Caso 3. Las varianzas poblacionales son diferentes: $\sigma_1^2 \neq \sigma_2^2$

Un intervalo de confianza para la diferencia de medias poblacionales $\mu_1 - \mu_2$, a un nivel del $100(1-\alpha)\,\%$, está dado por

$$\left((\overline{x}_1 - \overline{x}_2) - t_{\alpha/2}(g)\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}};\ (\overline{x}_1 - \overline{x}_2) + t_{\alpha/2}(g)\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right).$$

Donde $g = \dfrac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\dfrac{\left(s_1^2/n_1\right)^2}{n_1-1} + \dfrac{\left(s_2^2/n_2\right)^2}{n_2-1}}$.

Ejemplos

1. Se pretende comparar la duración de dos marcas de pilas alcalinas. Para ello se escogieron dos muestras de cinco pilas cada una. Los datos se presentan a continuación.

Marca A	100	96	92	96	92
Marca B	76	80	75	84	82

Si suponemos que las varianzas poblacionales son $\sigma_A^2 = 11$ y $\sigma_B^2 = 15$, determine, basándose en un intervalo de confianza de nivel 95 %, si las dos marcas de pilas tienen igual duración.

Solución: Se tiene que

$$n_1 = 5,\quad \overline{x}_1 = 95.2;\quad \sigma_1^2 = 11,$$
$$n_2 = 5,\quad \overline{x}_2 = 79.4;\quad \sigma_2^2 = 15.$$

Entonces,

$$\left((\overline{x}_1 - \overline{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}};\ (\overline{x}_1 - \overline{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right) =$$

$$\left((95.2-79.4) - 1.96\sqrt{\frac{11}{5}+\frac{15}{5}};\ (95.2-79.4) + 1.96\sqrt{\frac{11}{5}+\frac{15}{5}}\right) = (11.33; 20.27).$$

Si las dos medias fueran estadísticamente iguales, su diferencia sería igual a cero. Como cero no se encuentra en el intervalo de confianza, podemos afirmar que las dos marcas no tienen igual durabilidad.

2. Un ingeniero desea determinar si los automóviles americanos y los japoneses tienen igual consumo de combustible. Para ello escogío una muestra de 10 carros americanos y 12 japoneses de similares características y midió el consumo por 100 km de recorrido, con los siguientes resultados:

Americanos	6.0	5.8	7.4	7.6	8.5	9.2	7.6	6.7	10.0	9.8		
Japoneses	4.8	5.1	5.3	6.2	5.7	5.4	6.7	7.2	8.0	5.6	6.6	5.9

Estimar, mediante un intervalo de confianza al 95 %, la diferencia entre las dos medias de consumo. ¿Se puede decir el consumo de combustible no depende del origen del auto?

Solución: Supondremos que las poblaciones son normales con varianzas desconocidas, supuestas iguales. Además, se tienen los siguientes resultados:

$$n_1 = 10, \quad \overline{x}_1 = 7.86; \quad s_1^2 = 2.216,$$
$$n_2 = 12, \quad \overline{x}_2 = 6.04; \quad s_2^2 = 0.879.$$

El estimador de la varianza es

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{(10-1)2.216 + (12-1)0.879}{10+12-2} = 1.481.$$

Entonces, el intervalo es

$$\left((\overline{x}_1 - \overline{x}_2) - t_{\alpha/2}(n_1+n_2-2)\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}; (\overline{x}_1 - \overline{x}_2) + t_{\alpha/2}(n_1+n_2-2)\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}} \right) =$$

$$\left((7.86 - 6.04) - t_{0.025}(20)\sqrt{\frac{1.481}{10} + \frac{1.481}{12}}; (7.86 - 6.04) + t_{0.025}(20)\sqrt{\frac{1.481}{10} + \frac{1.481}{12}} \right) =$$
$$(0.733; 2.907).$$

Como el valor cero no se encuentra en el intervalo de la diferencia de medias, deducimos que la media de los consumos no es igual. Entonces, el consumo de combustible si depende del país de origen del carro.

7.15.2. Intervalo de confianza para la razón entre dos varianzas

Suponga que disponemos de dos muestras independientes de tamaños n_1 y n_2, seleccionadas de dos poblaciones que siguen leyes normales con varianzas σ_1^2 y σ_2^2, respectivamente. Deseamos construir un intervalo de confianza para la razón de las dos varianzas.

Un intervalo de confianza para la razón de las varianzas poblacionales $\frac{\sigma_1^2}{\sigma_2^2}$, a un nivel del $100(1-\alpha)$ %, está dado por

$$\left(\frac{s_1^2}{s_2^2} F_{1-\alpha/2}(n_1-1, n_2-1); \frac{s_1^2}{s_2^2} F_{\alpha/2}(n_1-1, n_2-1) \right).$$

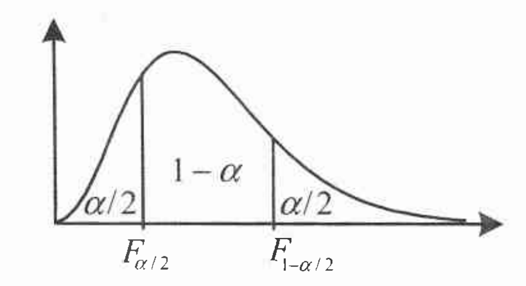


Figura 7.5: Localización de los valores de la ley F en el intervalo de confianza para la razón entre dos varianzas.

Ejemplo. Un inversionista quiere comparar, en términos de las varianzas, los rendimientos de las acciones de dos compañías del sector servicios. Calculó los rendimientos mensuales del último semestre, de las dos compañías, como se muestra a continuación.

Compañía A	1.8	0.3	2.4	0.1	1.5	1.9
Compañía B	2.6	3.8	3.7	4.1	3.2	3.5

Con el empleo de un intervalo de confianza de nivel 95 % para las varianzas, determine si los rendimientos tienen igual variabilidad.

Solución: Tenemos que

$$\begin{aligned} n_1 &= 6, \quad s_1^2 = 0.859,\\ n_2 &= 6, \quad s_2^2 = 0.278. \end{aligned}$$

El intervalo queda:

$$\begin{aligned} \left(\frac{s_1^2}{s_2^2}F_{1-\alpha/2}(n_1-1,n_2-1);\ \frac{s_1^2}{s_2^2}F_{\alpha/2}(n_1-1,n_2-1)\right) &= \\ \left(\frac{0.859}{0.278}F_{1-0.025}(5,5);\ \frac{0.859}{0.278}F_{0.025}(5,5)\right) &= (0.433; 22.082). \end{aligned}$$

Si las varianzas son iguales, su cociente es igual a uno. En este caso, el valor 1 se encuentra dentro del intervalo; por tanto, las varianzas de los rendimientos de las dos compañías son iguales.

7.15.3. Intervalo de confianza para la diferencia entre dos proporciones

Sean $\widehat{p}_1$ y $\widehat{p}_2$ las proporciones de éxitos de dos muestras aleatorias independientes, de tamaños n_1 y n_2, provenientes de dos poblaciones de Bernoulli, $\mathbf{Ber}(p_1)$ y $\mathbf{Ber}(p_2)$, respectivamente. Ahora, deseamos estimar, mediante un intervalo, la diferencia de esas proporciones poblacionales.

Un intervalo de confianza aproximado para la diferencia de proporciones $p_1 - p_2$, a un nivel de $100(1-\alpha)\,\%$ viene dado por

$$\left((\widehat{p}_1-\widehat{p}_2) - z_{\alpha/2}\sqrt{\frac{\widehat{p}_1\widehat{q}_1}{n_1}+\frac{\widehat{p}_2\widehat{q}_2}{n_2}};\ (\widehat{p}_1-\widehat{p}_2) + z_{\alpha/2}\sqrt{\frac{\widehat{p}_1\widehat{q}_1}{n_1}+\frac{\widehat{p}_2\widehat{q}_2}{n_2}}\right).$$

Donde $\widehat{q}_1 = 1-\widehat{p}_1$ y $\widehat{q}_2 = 1-\widehat{p}_2$.

Ejemplo. Un fabricante cree que su marca de yogurt es más preferida en la Sierra que en la Costa. Para probar su sospecha escogío dos muestras aleatorias de 500 consumidores en la Costa y 300 en

la Sierra. Las muestras revelaron que 210 consumidores en la Sierra y 320 en la Costa consumen su producto. De acuerdo a un intervalo de confianza al 98 %, ¿se puede inferir que el fabricante tiene razón?

Solución: Se tiene que

$$\textbf{Costa:}\quad n_1 = 500,\quad \widehat{p}_1 = \frac{320}{500} = 0.64,$$
$$\textbf{Sierra:}\quad n_2 = 300,\quad \widehat{p}_2 = \frac{210}{300} = 0.7.$$

Por tanto, el intervalo es

$$\left((\widehat{p}_1 - \widehat{p}_2) - z_{\alpha/2}\sqrt{\frac{\widehat{p}_1\widehat{q}_1}{n_1} + \frac{\widehat{p}_2\widehat{q}_2}{n_2}}; (\widehat{p}_1 - \widehat{p}_2) + z_{\alpha/2}\sqrt{\frac{\widehat{p}_1\widehat{q}_1}{n_1} + \frac{\widehat{p}_2\widehat{q}_2}{n_2}}\right) =$$
$$\left((0.64 - 0.7) - 2.33\sqrt{\frac{0.64\times 0.36}{500} + \frac{0.7\times 0.3}{300}}; (0.64 - 0.7) + 2.33\sqrt{\frac{0.64\times 0.36}{500} + \frac{0.7\times 0.3}{300}}\right) =$$
$$(-0.1394; 0.0194).$$

Como cero está dentro del intervalo, no se puede afirmar que la sospecha del fabricante sea cierta. ◄

En el Cuadro 7.2 se encuentra un resumen de los intervalos de confianza de dos muestras analizados en este capítulo.

	Hipótesis	Parámetro	Intervalo
Distribución general	σ_1 y σ_2 conocidas	$\mu_1 - \mu_2$	$(\overline{x}_1 - \overline{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
Distribución normal	σ_1 y σ_2 desconocidas supuestas iguales	$\mu_1 - \mu_2$	$(\overline{x}_1 - \overline{x}_2) \pm t_{\alpha/2;(n_1+n_2-2)}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
Distribución normal	σ_1 y σ_2 desconocidas distintas	$\mu_1 - \mu_2$	$(\overline{x}_1 - \overline{x}_2) \pm t_{\alpha/2;g}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
Distribución normal		$\frac{\sigma_1^2}{\sigma_2^2}$	$\left(\frac{s_1^2}{s_2^2}F_{\alpha/2;(n_1-1,n_2-1)}, \frac{s_1^2}{s_2^2}F_{1-\alpha/2;(n_1-1,n_2-1)}\right)$
Distribución binomial		$p_1 - p_2$	$(\widehat{p}_1 - \widehat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}_1\widehat{q}_1}{n_1} + \frac{\widehat{p}_2\widehat{q}_2}{n_2}}$

Cuadro 7.2: Intervalos de confianza sobre dos muestras comunmente empleados

A lo largo de este capítulo se ha enfatizado la idea de que los intervalos estadísticos reflejan la incertidumbre debida a la variabilidad de los datos, pero en la mayoría de casos prácticos, a más de que las hipótesis básicas sobre la muestra (como aleatoriedad, independencia y normalidad) son violadas, influyen otros factores cuya magnitud es incuantificable, por ejemplo, el comportamiento de las personas y los equipos, el medio ambiente, etc.

Entonces, en la correcta interpretación de los intervalos estadísticos habrá que tomar en cuenta que ellos no reflejan la incertidumbre total presente en las mediciones y solo proveen una *cota inferior* de la verdadera incertidumbre; por lo tanto, son una cruda aproximación a la realidad.

7.16. Ejercicios

1. El gerente de una empresa que tiene dos locales de ferretería cree que las ventas en el local del norte son mayores que en el local del sur. Para verificarlo, tomó una muestra de 200 facturas en el local del sur y 250 facturas del local del norte, resultando un promedio de 13 y 15 dólares y desviaciones estándar de 3 y 4 dólares, respectivamente. Mediante un intervalo de confianza de nivel 95 %, ¿se puede concluir que el gerente tiene razón?

2. Una organización de defensa de los derechos civiles afirma que en la industria de la construcción el salario medio semanal de los hombres supera en 13 dólares al de las mujeres. Una muestra aleatoria de 20 hombres y otra de 25 mujeres reveló promedios de 110 y 100 dólares, respectivamente. Si las dos poblaciones de salarios son normales con varianzas de 100 y 64, mediante un intervalo de confianza para la diferencia de medias, de nivel 98 %, ¿se puede decir que la afirmación es cierta?

3. Se cree que el precio de arriendo de las casas es mayor en Cuenca que en Ambato. Estudios anteriores revelan que las dos poblaciones de arriendos tienen distribución normal con varianza homogénea. Dos muestras aleatorias de tamaño 16 revelaron lo siguiente: $\overline{x}_1 = 138$, $s_1 = 6$ y $\overline{x}_2 = 135$, $s_2 = 4$. Con el empleo de un intervalo de confianza al 95 %, ¿se puede concluir que la suposición es correcta?

4. Un inversionista hizo un estudio para determinar en qué ciudad, Manta o Loja, podría abrir un supermercado. En una muestra de 21 hogares de la ciudad de Manta halló un ingreso promedio de 500 dólares y una desviación estándar de 120 dólares. En otra muestra de 16 hogares de la ciudad de Loja halló un ingreso promedio de 480 dólares y desviación estándar de 60 dólares. Suponiendo poblaciones normales con varianzas diferentes y con el empleo de un intervalo de nivel 95 %, ¿en cuál de las ciudades le conviene abrir el supermercado?

5. El dueño de dos restaurantes que vende pollo a la brasa quiere determinar si sus dos locales venden la misma cantidad de pollos diarios. Dos muestras de las ventas de 12 días dieron los siguientes números de pollos vendidos:

Local A	12	17	14	18	9	10	20	15	12	16	8	14
Local B	12	14	13	11	12	15	21	14	13	14	22	15

   Además, las muestras revelaron que las dos poblaciones de muestras son normales con varianzas diferentes. ¿Se puede inferir que en los dos locales las ventas diarias de pollos son las mismas?

6. Una manera de comparar el riesgo de dos inversiones es a través de sus varianzas. Para una inversión en la industria electrónica se tomó una muestra de 10 datos y resultó una desviación estándar de 52. Para una inversión en la industria siderúrgica se tomó una muestra de 15 datos y resultó una desviación estándar de 31. Si se asume que cada una de las inversiones siguen leyes normales, ¿cuál es su conclusión si utiliza un intervalo de confianza al 90 % para la razón de varianzas?

7. Un investigador sospecha que los hombres y las mujeres difieren significativamente en tiempo diario de utilización del teléfono. Entrevista a 25 sujetos de cada sexo obteniendo los siguientes resultados:

   Mujeres: $\overline{x}_1 = 38;\quad \sigma_1 = 6,$
   Hombres: $\overline{x}_2 = 31;\quad \sigma_2 = 5.$

   Utilice un nivel de confianza del 95 % para:

   a) construir un intervalo para la razón entre las varianzas. ¿Se puede decir que son iguales?;

b) construir un intervalo para la diferencia de medias adecuado.

8. El entrenador de un equipo de fútbol quiere comparar la efectividad de sus jugadores al cobrar un tiro penal. En los entrenamientos, los jugadores nacionales, de 120 tiros intentados, acertaron 91, mientras que los extranjeros de 90 tiros realizados, acertaron 78. Mediante un intervalo de confianza de nivel 90 %, ¿se puede inferir que no hay diferencia entre los jugadores nacionales y extranjeros al cobrar un tiro penal?

9. En un estudio epidemiológico, se tomó una muestra aleatoria de 300 hombres y 27 de ellos padecieron o padecen una determinada variedad de gripe. También, se tomó una muestra de 400 mujeres y a 32 les ocurría lo mismo. A la vista de estos datos, ¿se puede considerar que este tipo de gripe afecta a hombres y mujeres por igual? (Use $1 - \alpha = 0.96$)

10. Se quiere comprobar la efectividad de una vacuna contra una enfermedad y para ello, tras contagiar a 200 animales, se la suministra a 100 y se compara con otros 100, a los cuales no se les suministró. De entre los vacunados, mueren 8 a causa de la enfermedad y de los no vacunados 20. A un nivel de confiabilidad del 93 %, ¿podemos decir que la vacuna es eficaz para reducir la tasa de mortalidad?

Capítulo 8

Pruebas de Hipótesis

Dudar sobre todo o creer en todo: estas son dos estrategias igualmente convenientes. Con ambas nos evitamos la necesidad de reflexionar
Henry Poncairé

En el capítulo anterior se vio que se puede realizar inferencias acerca de un parámetro poblacional estimando su valor, ya sea de manera puntual o como un intervalo de confianza. Pero en muchas ocasiones no interesa conocer o tener una idea del valor del parámetro, sino comprobar (o rechazar) una afirmación sobre el valor que tiene el parámetro, sin importar la longitud o la localización del intervalo.

Supongamos que un investigador desea probar que actualmente, el ingreso mensual de los ecuatorianos es mayor que el ingreso que tenían 5 años atrás. Para verificarlo, recolecta los datos mediante una muestra tomada al azar. El investigador podría desear comprobar la hipótesis de que el aumento del ingreso es mayor que un cierto valor prefijado. Un intervalo de confianza de la variación media del ingreso proporcionará menor información que una prueba de hipótesis sobre la variación del ingreso.

Las hipótesis son, en general, afirmaciones sobre los parámetros poblacionales, como la media y la varianza; así, se pudiera desear probar que el ingreso medio actual no es diferente de aquel que los ecuatorianos tenían hace 5 años. Una hipótesis también puede ser una afirmación sobre la distribución de una característica de interés; por ejemplo, que el ingreso mensual sigue una distribución normal.

La teoría de las pruebas de hipótesis sobre parámetros poblacionales fue desarrollada en los primeros años del siglo XX y sistematizada por R. A. Fisher, E. S. Pearson y J. Neyman. Éstos últimos la formalizaron e introdujeron el vocabulario actualmente en uso, en una publicación realizada en 1933. Aquí se expondrán los casos de mayor uso e importancia.

8.1. Elementos de una prueba de hipótesis

Para probar una hipótesis estadística es necesario tener en cuenta ciertos elementos que conducirán a aceptar o rechazar la hipótesis planteada, de manera correcta. Ello lo ilustraremos con el ejemplo anterior.

El investigador cree que, en condiciones normales, el aumento del ingreso medio θ debe ser menor que un nivel prefijado $\theta_0 = 60$ dólares, y decide probar esta afirmación; pero en el curso de la investigación

se podría concluir que la variación del ingreso es mayor que 60 dólares. Para tomar cualquier decisión es necesario que el investigador compare con un valor que le informe de la validez o invalidez de su hipótesis.

En la terminología de las pruebas de hipótesis, aquella que especifica un valor particular del parámetro que se estudia se llama *hipótesis nula*, que se representa por H_0. Esta hipótesis, usualmente, corresponde al procedimiento de operación normal de un sistema de especificaciones conocidas. En el ejemplo, $\theta \leq 60$ es la hipótesis nula, pues representa lo que debería suceder cuando ha habido un incremento del ingreso, en condiciones normales; o sea, H_0: $\theta \leq 60$.

La hipótesis que especifica aquellos valores del parámetro que representan un cambio importante del procedimiento normal o de las especificaciones conocidas, se llama *hipótesis alternativa* y se representa por H_1. En el ejemplo los valores mayores que 60 indicarían un comportamiento anómalo o extraordinario; así, la hipótesis alternativa es H_1: $\theta > 60$.

La cantidad, calculada a partir de la muestra, que permite decidir si la hipótesis nula será o no será rechazada se denomina *estadístico de prueba*. La distribución de probabilidad del mencionado estadístico debe ser conocida para poder realizar la prueba.

El conjunto de valores del estadístico de prueba que conduce al rechazo de la hipótesis nula, en favor de la hipótesis alternativa, se llama *región de rechazo* o región crítica de la prueba. La *decisión* consiste en rechazar la hipótesis nula en favor de la alternativa si el valor del estadístico de prueba cae en la región de rechazo; caso contrario no se descarta H_0.

En las pruebas de hipótesis, el resultado se da en términos de la hipótesis nula. Nosotros o «rechazamos H_0 en favor de H_1» o «no rechazamos H_0»; nunca concluimos «rechazamos H_1» o «aceptamos H_1». Luego, y en base del análisis del valor del estadístico de prueba, se podría sacar otras conclusiones sobre H_1, pero ésto dependerá de cada caso particular.

A continuación se resumen los elementos de una prueba de hipótesis:

1. Hipótesis nula, H_0.
2. Hipótesis alternativa, H_1.
3. Estadístico de prueba.
4. Región de rechazo.
5. Decisión.

Errores de tipo I y de tipo II

Al tomar una decisión se puede cometer dos tipos de errores: rechazar la hipótesis nula cuando es verdadera, lo que se llama error de tipo I; o bien, no rechazar la hipótesis nula cuando en realidad es falsa, lo que se denomina error de tipo II. Ello se resume en el siguiente cuadro:

Decisión	**Hipótesis Nula**	
	Verdadera	Falsa
Rechazar H_0	Error tipo I	Decisión correcta
No rechazar H_0	Decisión correcta	Error tipo II

Entonces, tenemos las siguiente definiciones:

1. La probabilidad de cometer un error de tipo I se denota por α y se fija al escoger la región de rechazo. Este valor se denomina *nivel de significación* de la prueba estadística:

$$\Pr(\text{error de tipo I}) = \text{nivel de significación} = \alpha.$$

2. El *nivel de significación observado* o $p-$valor, es el mínimo valor del nivel de significación para el cual los datos observados indican que la hipótesis nula debe ser rechazada.

   Este valor se utiliza cuando se trabaja con paquetes computacionales para determinar si se rechaza H_0. La regla es la siguiente:

Si el valor de $p < \alpha$, entonces se rechaza H_0. Caso contrario, no se rechaza (se acepta).

3. La probabilidad de cometer un error de tipo II se denota por β y depende de la hipótesis alternativa que se escoja:

$$\Pr(\text{error de tipo II}) = \beta.$$

4. La *potencia* de una prueba es la probabilidad de tomar la decisión acertada, de rechazar H_0 cuando ésta es falsa, y es igual a $1 - \beta$.

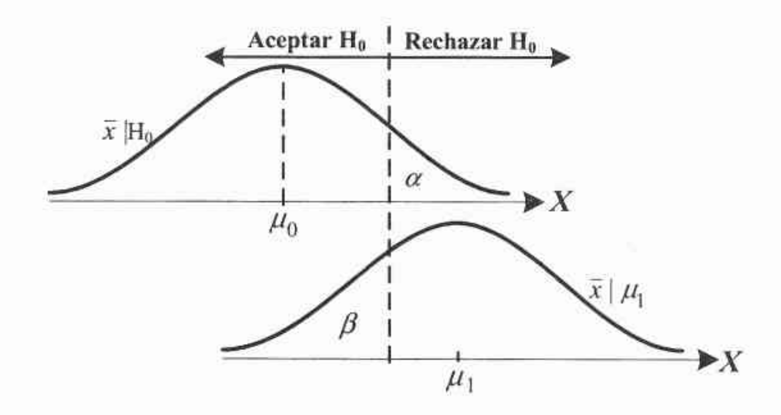


Figura 8.1: Probabilidad de cada tipo de error al realizar una prueba.

Las probabilidades α y β tienen las siguientes propiedades:

1. Para un tamaño de muestra fijo, al aumentar la región de rechazo (y por lo tanto α), β disminuye.

2. Al aumentar el tamaño muestral n, las probabilidades α y β decrecen a la vez.

Un error de tipo I es considerado más serio, y por lo tanto más importante de evitar, que un error de tipo II. Consecuentemente, el procedimiento de probar una hipótesis se ajusta de manera que se garantice una «baja» probabilidad de rechazar erroneamente la hipótesis nula.

Frecuentemente, se determina el nivel α de manera arbitraria, pero si no se lo especifica, habrá que calcular el $p-$valor del contraste. Esto permite rechazar H_0 para valores pequeños de α. Nosotros tomaremos, de preferencia, valores para α de 0.1, 0.05 o 0.01.

Tipos de pruebas de hipótesis

Las pruebas de hipótesis se clasifican en unilaterales y bilaterales.

Definición (de prueba estadística unilateral) Una prueba estadística unilateral es aquella en la que la región de rechazo se localiza solamente en un extremo de la distribución de probabilidad del estadístico de prueba.

Para detectar si $\theta \leq \theta_0$ la región de rechazo se sitúa en el extremo superior de la distribución del estimador $\widehat{\theta}$ (Figura 8.2). Para detectar si $\theta \geq \theta_0$ la región de rechazo se sitúa en el extremo inferior de la distribución de $\widehat{\theta}$.

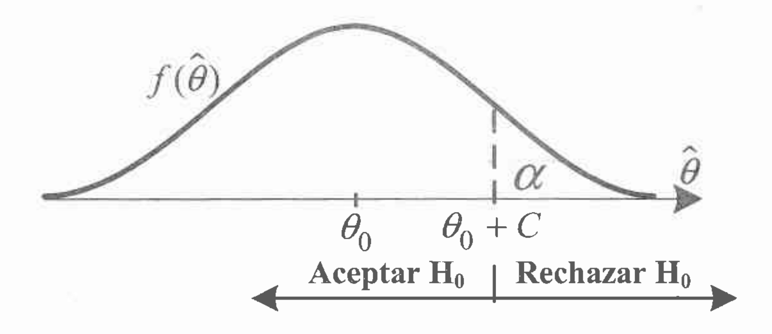


Figura 8.2: Distribución de $\widehat{\theta}$ cuando se cumple H_0: $\theta \leq \theta_0$.

Definición (de prueba estadística bilateral) Una prueba estadística bilateral es aquella que sitúa la región de rechazo en ambos extremos de la distribución de probabilidad del estadístico de prueba.

Las pruebas bilaterales se utilizan para detectar $\theta < \theta_0$ o bien $\theta > \theta_0$; es decir, $\theta \neq \theta_0$ (Figura 8.3).

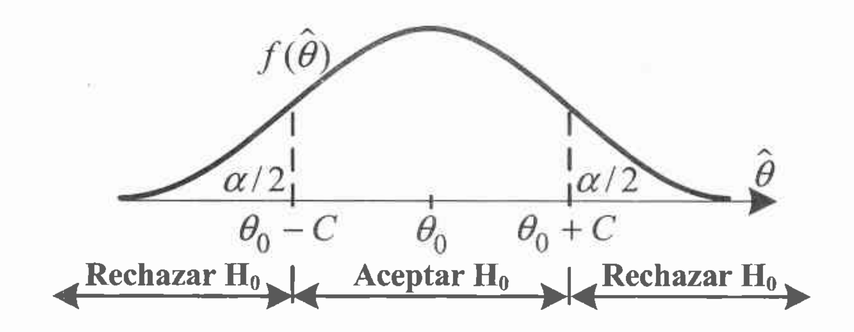


Figura 8.3: Distribución de $\widehat{\theta}$ cuando se cumple H_0: $\theta = \theta_0$.

8.2. Pruebas de hipótesis sobre la media, cuando la varianza es conocida

Uno de los casos más comunes en la realización de pruebas de hipótesis es hacerla sobre el valor de la media poblacional, cuando se dispone de una muestra de dicha población. El parámetro θ que se desea probar es μ y el estimador $\widehat{\theta}$ es la media muestral $\overline{x}$.

A continuación se exponen –de manera sucinta–, las pruebas estadísticas, bilateral y unilaterales, cuando deseamos probar que el valor de la media poblacional μ es igual a un valor prefijado μ_0.

a) **Prueba bilateral para la media.**

1. *Hipótesis Nula.* H_0: $\mu = \mu_0$.

2. *Hipótesis Alternativa.* H_1: $\mu \neq \mu_0$.
3. *Estadístico de Prueba.* $z_{obs} = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$.
4. *Región de Rechazo.* $z_{obs} < -z_{\alpha/2}$ o $z_{obs} > z_{\alpha/2}$.

Ejemplo. Una empresa farmacéutica ha establecido que un comprimido debe tener un peso medio igual a $\mu_0 = 0.5\,\text{g}$ y una desviación estándar de $\sigma = 0.11\,\text{g}$. Se tomó una muestra de 144 comprimidos de un lote de fármacos, cuyo peso promedio fue de $\overline{x} = 0.53\,\text{g}$.

a) Para un nivel de significación de 0.01, ¿el peso de los comprimidos en el lote se diferencia del admisible por la empresa?;

b) Determinar el $p-$valor de la prueba.

Solución:

a) La prueba de hipótesis es bilateral:

1. *Hipótesis Nula.* H_0: $\mu = 0.5\,\text{g}$.
2. *Hipótesis Alternativa.* H_1: $\mu \neq 0.5\,\text{g}$.
3. *Estadístico de Prueba.* $z_{obs} = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \dfrac{0.53 - 0.5}{0.11/\sqrt{144}} = 3.273$.
4. *Región de Rechazo.* Para $\alpha = 0.01$, la región de rechazo es $z_{obs} < -2.57$ o $z_{obs} > 2.57$.
5. *Decisión.* Como el valor z_{obs} cae en la región de rechazo, se descarta la hipótesis de que $\mu = 0.5$. La probabilidad de rechazar H_0, suponiendo que sea cierta es solo 0.01; por lo tanto, en al menos el 99 % de las veces la decisión es la correcta.

b) Determinemos el nivel de significación mínimo de la prueba.
Como $z_{obs} = 3.273$, el valor de probabilidad correspondiente es $\Phi(3.273) = 0.9995$. Por ser una prueba bilateral, se cumple que $0.9995 + \dfrac{\alpha}{2} = 1$; por lo tanto, $\alpha = 2(1 - 0.9995) = 0.001$; que significa que llegaremos a una conclusión errónea en menos de 1 de cada 1000 veces, lo que verifica la conclusión anterior. ◀

b) **Prueba unilateral para la media.**

1. *Hipótesis Nula.* H_0: $\mu = \mu_0$.
2. *Hipótesis Alternativa.* H_1: $\mu > \mu_0$ (o bien H_1: $\mu < \mu_0$).
3. *Estadístico de Prueba.* $z_{obs} = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$.
4. *Región de Rechazo.* $z_{obs} > z_\alpha$ (o bien $z_{obs} < -z_\alpha$, cuando H_1: $\mu < \mu_0$).

Ejemplo. Si en el ejemplo anterior, el peso máximo admisible para que el medicamento no sea tóxico es igual a $\mu_0 = 0.52\,\text{g}$.

a) Se desea saber si los comprimidos del lote son aptos para el consumo humano, a un nivel de significación del 5 %;

b) Determinar el nivel de significación de la prueba.

Solución:

a) Los datos son los mismos que antes se usaron, solo debiéndose cambiar las hipótesis.

1. *Hipótesis Nula.* H_0: $\mu = 0.52\,\text{g}$.
2. *Hipótesis Alternativa.* H_1: $\mu > 0.52\,\text{g}$.
3. *Estadístico de Prueba.* $z_{obs} = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \dfrac{0.53 - 0.52}{0.11/\sqrt{144}} = 1.0909.$
4. *Región de Rechazo.* Para $\alpha = 0.05$ la región de rechazo es $z_{obs} > 1.65$.
5. *Decisión.* Como el valor z_{obs} no cae en la región crítica, no existe razón para descartar la hipótesis nula, de manera que se podría asegurar que el medicamento es apto para el consumo humano.

b) El nivel de significación de la prueba es $1 - \Phi(1.0909) = 1 - 0.8621 = 0.1379$. Como $p = 0.1379 > 0.05$, no se rechaza H_0. ◄

Observación. Si el tamaño de la muestra es suficientemente grande ($n \geq 25$) y se desconoce la varianza, es posible utilizar estas pruebas, reemplazando σ por su estimador s, sin pérdida de exactitud.

Cálculo de la potencia de la prueba

La probabilidad de cometer un error de tipo II se nota como β y se define por

$$\beta = \Pr(\text{error de tipo II}) = \Pr(\text{aceptar } H_0|\mu_1),$$

donde suponemos verdadera la hipótesis H_1: $\mu = \mu_1$.

Entonces, por el Teorema del Límite Central, la variable $Z = \dfrac{\overline{x} - \mu_1}{\sigma/\sqrt{n}}$ sigue una ley normal estándar.

Ejemplo. En el ejemplo anterior, calcular la potencia de la prueba si el verdadero valor de la media es 0.54 mg.

Solución: Veamos qué significa «aceptar H_0» en términos de los valores que puede tomar el promedio, $\overline{x}$.

Aceptamos H_0 si $\dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \leq 1.65$; es decir, $\dfrac{\overline{x} - 0.52}{0.11/\sqrt{144}} \leq 1.65$, o sea $\overline{x} \leq 0.53513$.

Por tanto, la probabilidad β la podemos poner de la siguiente manera:

$$\begin{aligned}
\Pr(\text{aceptar } H_0|\mu_1) &= \Pr(\overline{x} \leq 0.53513|\mu_1 = 0.54) \\
&= \Pr\left(\frac{\overline{x} - \mu_1}{\sigma/\sqrt{n}} \leq \frac{0.53513 - 0.54}{0.11/\sqrt{144}}\right) \\
&= \Pr(Z \leq -0.53127) = 0.2976.
\end{aligned}$$

La correspondiente potencia de la prueba es **Pot** $= 1 - \beta = 0.7024$.

8.3. Pruebas de hipótesis sobre la media, cuando la varianza es desconocida

Cuando la varianza es desconocida, no es posible aplicar el Teorema del Límite Central; en este caso, para que sea posible aplicar esta prueba es necesario que la muestra provenga de una población que sigue una ley normal, de manera que el estadístico de prueba sigue una ley de distribución t. Entonces, la prueba estadística es la siguiente:

a) **Prueba bilateral para la media.**

1. *Hipótesis Nula.* H_0: $\mu = \mu_0$.
2. *Hipótesis Alternativa.* H_1: $\mu \neq \mu_0$.
3. *Estadístico de Prueba.* $t_{obs} = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$.
4. *Región de Rechazo.* $t_{obs} < -t_{\alpha/2}(n-1)$ o $t_{obs} > t_{\alpha/2}(n-1)$.

Ejemplo. Según un estudio del Ministerio de Educación, el costo medio de la lista de útiles de los escolares de educación básica es 87 dólares. Para verificarlo, un investigador tomó una muestra con los siguientes resultados:

Costo (x_i)	68	75	93	101	123
Número (n_i)	2	3	4	6	5

Para un nivel de significación de 0.05, verificar la hipótesis de que la máquina cumple con la especificación.

Solución: Previamente hallamos el promedio y la desviación estándar: $\overline{x} = 97.7$ y $s = 18.728$ (los cálculos se dejan como ejercicio para el lector).

Con ésto planteamos el contraste:

1. *Hipótesis Nula.* H_0: $\mu = 87$.
2. *Hipótesis Alternativa.* H_1: $\mu \neq 87$.
3. *Estadístico de Prueba.* $t_{obs} = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}} = \dfrac{97.7 - 87}{18.728/\sqrt{20}} = 2.555$.
4. *Región de Rechazo.* Para el nivel de significación $\alpha = 0.05$ y por el número de grados de libertad $(n-1) = 19$ se encuentra el valor de $t_{0.025}(19) = 2.093$. La región de rechazo es $t_{obs} < -2.093$ o $t_{obs} > 2.093$.
5. *Decisión.* Como el valor t_{obs} se encuentra en la región crítica, ya que $2.555 > 2.093$, se rechaza la hipótesis nula; es decir, el precio medio de las listas de útiles es distinto al que afirma el Ministerio.

b) **Prueba unilateral de la media.**

1. *Hipótesis Nula.* H_0: $\mu = \mu_0$.
2. *Hipótesis Alternativa.* H_1: $\mu > \mu_0$ (o bien H_1: $\mu < \mu_0$).
3. *Estadístico de Prueba.* $t_{obs} = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$.
4. *Región de Rechazo.* $t_{obs} > t_\alpha(n-1)$ (o bien $t_{obs} < -t_\alpha(n-1)$, cuando H_1: $\mu < \mu_0$).

Ejemplo. Según las previsiones del gobierno, la inflación para este año será de 3.9 %. Un economista, desconfiado de la cifra, realizó una investigación por su cuenta y registró la variación de los precios en los 22 artículos que a su juicio tienen la mayor incidencia en la economía popular. Obtuvo una variación de 4.5 % y una desviación estándar de 1.3 %. Pruebe si la cifra de inflación del investigador será mayor que la del gobierno.

Solución: Se tienen los siguientes datos del problema: $n = 22$, $\overline{x} = 4.5$ y $s = 1.3$. La prueba es:

1. *Hipótesis Nula.* H_0: $\mu = 3.9$.
2. *Hipótesis Alternativa.* H_1: $\mu > 3.9$.

3. *Estadístico de Prueba.* $t_{obs} = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}} = \dfrac{4.5 - 3.9}{1.3/\sqrt{22}} = 2.165.$
4. *Región de Rechazo.* Si tomamos $\alpha = 0.01$ y $t_{0.01}(21) = 2.518$, la región crítica es $t_{obs} > 2.518$; pero si se toma $\alpha = 0.05$ y $t_{0.05}(21) = 1.721$, la región crítica es $t_{obs} > 1.721$.
5. *Decisión.* Para $\alpha = 0.01$ no se rechaza H_0; es decir, no existe evidencia que indique que la cifra dada por el gobierno está subestimada. Sin embargo, si se toma $\alpha = 0.05$, resulta que $t_{obs} > t_\alpha$, se rechaza H_0; es decir, la inflación es mayor que la estimada por el gobierno.

A partir de los dos resultados se deduce que la aceptación o rechazo de $\mu = 3.9$ no es concluyente, puesto que con $\alpha = 0.05$ se rechaza y con $\alpha = 0.01$ se acepta. En este caso se debería aumentar el número de observaciones.

8.4. Pruebas de hipótesis sobre la varianza

Para realizar una prueba de hipótesis sobre la varianza, supondremos que las observaciones provienen de una distribución normal, para que el estadístico $\dfrac{(n-1)s^2}{\sigma^2}$ siga una distribución χ^2 con $(n-1)$ grados de libertad. Bajo este supuesto, las pruebas de hipótesis son las siguientes.

a) **Prueba bilateral para la varianza.**

1. *Hipótesis Nula.* H_0: $\sigma^2 = \sigma_0^2$.
2. *Hipótesis Alternativa.* H_1: $\sigma^2 \neq \sigma_0^2$.
3. *Estadístico de Prueba.* $\chi^2_{obs} = \dfrac{(n-1)s^2}{\sigma_0^2}$.
4. *Región de Rechazo.* $\chi^2_{obs} > \chi^2_{\alpha/2}(n-1)$ o $\chi^2_{obs} < \chi^2_{1-\alpha/2}(n-1)$.

b) **Prueba unilateral para la varianza.**

1. *Hipótesis Nula.* H_0: $\sigma^2 = \sigma_0^2$.
2. *Hipótesis Alternativa.* H_1: $\sigma^2 > \sigma_0^2$ (o bien H_1: $\sigma^2 < \sigma_0^2$).
3. *Estadístico de Prueba.* $\chi^2_{obs} = \dfrac{(n-1)s^2}{\sigma_0^2}$.
4. *Región de Rechazo.* $\chi^2_{obs} > \chi^2_{\alpha}(n-1)$ (o bien $\chi^2_{obs} < \chi^2_{1-\alpha}(n-1)$, cuando H_1: $\sigma^2 < \sigma_0^2$).

Ejemplo. Un fabricante de cables de cobre afirmó que su producto tenía una resistencia a la ruptura relativamente estable y que se ubicaría en un rango de 40 kilogramos-fuerza (kgf). Una muestra de 16 mediciones arrojó una varianza igual a $s^2 = 195$.

a) ¿Hay evidencia suficiente para rechazar la afirmación del fabricante?;

b) Encontrar el $p-$valor de la prueba.

Solución: Como el fabricante da el rango de variación de la resistencia, podemos estimar la desviación estándar mediante la relación aproximada $\sigma \approx \dfrac{\text{Rango}}{4}$ (dada en el Capítulo 1), por lo que se puede asumir que $\sigma = 10$ kgf.

a) La prueba es:

1. *Hipótesis Nula.* H_0: $\sigma^2 = 10^2 = 100$.
2. *Hipótesis Alternativa.* H_1: $\sigma^2 > 100$.
3. *Estadístico de Prueba.* $\chi^2_{obs} = \dfrac{(n-1)s^2}{\sigma_0^2} = \dfrac{(16-1)\times 195}{100} = 29.25$.
4. *Región de Rechazo.* Para un nivel de significación $\alpha = 0.05$ y 15 g.l., $\chi^2_{0.05}(15) = 25.00$. La región crítica es $\chi^2_{obs} > 25$.
5. *Decisión.* Como $29.25 > 25$, se concluye que la hipótesis es falsa y que la variación de las mediciones excede las especificaciones del fabricante.

b) Para encontrar el nivel de significación aproximado de la prueba examinamos, en la tabla de la distribución χ^2, el renglón correspondiente a 15 g.l. Se ve que el valor de $\chi^2_{obs} = 29.25$ es mayor que $\chi^2_{0.025} = 27.49$ y menor que $\chi^2_{0.01} = 30.58$. Así, el nivel de significación observado es menor que 0.025, lo que quiere decir que se rechazará la hipótesis nula para cualquier α mayor o igual al 2.5 %.

El valor exacto, obtenido mediante un programa informático, es 1.45 %.

8.5. Pruebas de hipótesis sobre la proporción

Supongamos que se dispone de n observaciones provenientes de una población con distribución de Bernoulli y deseamos probar que el parámetro p es igual a un valor prefijado p_0. Recordemos, también, que si entre las n observaciones hay y éxitos, la proporción se estima por $\widehat{p} = \dfrac{y}{n}$.

Para la realización de estas pruebas, utilizaremos la aproximación de la ley binomial mediante la ley normal.

a) **Prueba bilateral para la proporción.**

1. *Hipótesis Nula.* H_0: $p = p_0$.
2. *Hipótesis Alternativa.* H_1: $p \neq p_0$.
3. *Estadístico de Prueba.* $z_{obs} = \dfrac{\widehat{p} - p_0}{\sigma_{\widehat{p}}} = \dfrac{\widehat{p} - p_0}{\sqrt{\dfrac{p_0\, q_0}{n}}}$, donde $\widehat{p} = \dfrac{y}{n}$.
4. *Región de Rechazo.* $z_{obs} < -z_{\alpha/2}$ o $z_{obs} > z_{\alpha/2}$.

b) **Prueba unilateral para la proporción.**

1. *Hipótesis Nula.* H_0: $p = p_0$.
2. *Hipótesis Alternativa.* H_1: $p > p_0$ (o bien H_1: $p < p_0$).
3. *Estadístico de Prueba.* $z_{obs} = \dfrac{\widehat{p} - p_0}{\sqrt{\dfrac{p_0\, q_0}{n}}}$.
4. *Región de Rechazo.* $z_{obs} > z_\alpha$ (o bien $z_{obs} < -z_\alpha$, cuando H_1: $p < p_0$).

Ejemplo. Una empresa realizó una investigación de mercado para determinar el nivel de consumo de un refresco, para lo que consultó a 200 consumidores, de los cuales 28 expresaron su preferencia por el producto. El fabricante, de acuerdo a sus ventas, cree que tiene el 10 % del mercado de refrescos.

a) ¿Son los resultados de la investigación consistentes con los datos del fabricante?;

b) Determinar el nivel de significación del contraste.

Solución: Se tiene $p_0 = 0.1$ y el estimador de la proporción es $\widehat{p} = \dfrac{28}{200} = 0.14$.

a) La prueba queda así:

1. *Hipótesis Nula.* H_0: $p = 0.1$.
2. *Hipótesis Alternativa.* H_1: $p \neq 0.1$.
3. *Estadístico de Prueba.* $z_{obs} = \dfrac{\widehat{p} - p_0}{\sqrt{\dfrac{p_0 q_0}{n}}} = \dfrac{0.14 - 0.1}{\sqrt{\dfrac{0.1 \times 0.9}{200}}} = 1.886.$
4. *Región de Rechazo.* Al escoger $\alpha = 0.05$, $z_{0.025} = 1.96$, la región es $z_{obs} > 1.96$.
5. *Decisión.* Como z_{obs} no cae en la región de rechazo, no se puede descartar H_0; entonces, no hay evidencia de que la proporción de consumidores sea distinta del 10 %.

b) Como $z_{obs} = 1.886$, el valor de probabilidad correspondiente es $\Phi(1.886) = 0.9706$. Por ser una prueba bilateral, se cumple que $0.9706 + \dfrac{\alpha}{2} = 1$; por lo tanto, $\alpha = 0.0588$.

Cálculo de la potencia de la prueba

La probabilidad de cometer un error de tipo II se nota como β y se define por

$$\beta = \Pr(\text{error de tipo II}) = \Pr(\text{aceptar } H_0|p_1),$$

donde suponemos verdadera la hipótesis H_1: $p = p_1$. Entonces, la variable $Z = \dfrac{\widehat{p} - p_1}{\sqrt{\dfrac{p_1(1 - p_1)}{n}}}$ sigue una ley normal estándar.

Ejemplo. En el ejemplo anterior, calcular la potencia de la prueba si el verdadero valor de la proporción es 0.12.

Solución: Veamos qué significa «aceptar H_0» en función de $\widehat{p}$.

Aceptamos H_0 si $\left|\dfrac{\widehat{p} - p_0}{\sqrt{\dfrac{p_0 q_0}{n}}}\right| \leq 1.96$; es decir, $\left|\dfrac{\widehat{p} - 0.1}{\sqrt{\dfrac{0.1 \times 0.9}{200}}}\right| \leq 1.96$, o sea $0.05842 \leq \widehat{p} \leq 0.14159$.

Por tanto, la probabilidad β la podemos poner de la siguiente manera:

$$\begin{aligned}\Pr(\text{aceptar } H_0|p_1) &= \Pr(0.05842 \leq \widehat{p} \leq 0.14159|p_1 = 0.12) \\ &= \Phi\left(\frac{0.14159 - 0.12}{\sqrt{\frac{0.12 \times 0.88}{200}}}\right) - \Phi\left(\frac{0.05842 - 0.12}{\sqrt{\frac{0.12 \times 0.88}{200}}}\right) \\ &= \Phi(0.93958) - \Phi(-2.6799) = 0.8226.\end{aligned}$$

La correspondiente potencia de la prueba es **Pot** $= 1 - \beta = 0.1774$. ◄

En el Cuadro 8.1 se presenta un resumen de las pruebas de hipótesis con muestra única.

	Hipótesis nula (H_0)	Hipótesis alternativa (H_1)	Estadístico de prueba	Región de rechazo
Distribución general varianza conocida (muestra grande)	$\mu = \mu_0$ $\mu \leq \mu_0$ $\mu \geq \mu_0$	$\mu \neq \mu_0$ $\mu > \mu_0$ $\mu < \mu_0$	$z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$	$\lvert z\rvert > z_{\alpha/2}$ $z > z_\alpha$ $z < -z_\alpha$
Distribución normal varianza desconocida	$\mu = \mu_0$ $\mu \leq \mu_0$ $\mu \geq \mu_0$	$\mu \neq \mu_0$ $\mu > \mu_0$ $\mu < \mu_0$	$t = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$ $(n-1)$ g.l.	$\lvert t\rvert > t_{\alpha/2}$ $t > t_\alpha$ $t < -t_\alpha$
Distribución normal	$\sigma^2 = \sigma_0^2$ $\sigma^2 \leq \sigma_0^2$ $\sigma^2 \geq \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$ $\sigma^2 > \sigma_0^2$ $\sigma^2 < \sigma_0^2$	$\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$ $(n-1)$ g.l.	$\chi^2 < \chi^2_{1-\alpha/2}, \chi^2 > \chi^2_{\alpha/2}$ $\chi^2 > \chi^2_\alpha$ $\chi^2 < \chi^2_{1-\alpha}$
Distribución binomial (muestra grande)	$p = p_0$ $p \leq p_0$ $p \geq p_0$	$p \neq p_0$ $p > p_0$ $p < p_0$	$z = \dfrac{\widehat{p} - p_0}{\sqrt{\dfrac{p_0 q_0}{n}}}$	$\lvert z\rvert > z_{\alpha/2}$ $z > z_\alpha$ $z < -z_\alpha$

Cuadro 8.1: Pruebas de hipótesis con una sola muestra.

8.6. Ejercicios

Pruebas sobre la media poblacional (varianza conocida)

1. Se sospecha que los varones de las nuevas generaciones tienen, en promedio, mayor estatura que las antiguas. En un estudio realizado hace dos décadas se determinó que la población adulta masculina tenía una estatura media de 167 cm, con desviación estándar de 10 cm.

   a) Si se desea verificar la suposición anterior, formule, en símbolos y en palabras, las hipótesis nula y alternativa;

   b) Recientemente se tomó una muestra de 35 reclutas del servicio militar y se observó una estatura promedio de 171 cm. ¿Qué conclusión se puede sacar con $\alpha = 0.01$?

2. Un fabricante de fertilizantes afirma que el uso de su producto dará por resultado una cosecha de por lo menos 3.5 toneladas de trigo por hectárea, como media, con una desviación estándar de 0.5 toneladas. La aplicación del fertilizante a un área de 37 hectáreas dio una cosecha de 3.35 toneladas por hectárea.

   a) Al nivel de significación del 5 %, ¿se rechaza la afirmación del fabricante?;

   b) ¿Se rechaza la afirmación al nivel del 1 %?

3. Una balanza se encuentra descalibrada y no siempre registra el peso exacto. Cuando se pesan 454 g la desviación estándar es de 10 g. Para averiguar si es necesario recalibrar la balanza se realizó una serie de 50 pesajes de cantidades iguales de «una libra», resultando un peso promedio de 451 g.

   a) A un nivel del 10 %, ¿es necesario recalibrar la balanza?;

   b) Calcule el nivel de significación de la prueba.

4. Una empresa que elabora planchas plásticas puso en práctica un nuevo método de fabricación tal que el costo medio por metro cuadrado se distribuye normalmente con varianza poblacional 4. Se obtuvo una muestra aleatoria simple de tamaño 100, resultando un costo promedio de 2.4 dólares. Verifique la hipótesis de que el costo medio es de 3 dólares, con un nivel de significación del 4.5 %.

5. Por estudios anteriores, se sabe que la media de la edad de los delincuentes en el país es de 23 años; sin embargo, en un estudio reciente con una muestra de 80 detenidos, se obtuvo un promedio de 21.5 años y una desviación estándar de 3. Con un nivel de significación del 2 %, compruebe la hipótesis de que la media de edad de los delincuentes se ha reducido significativamente.

6. Un laboratorio farmacéutico asegura que un medicamento que fabrica tiene un contenido medio de 10 000 unidades en cada dosis. Para comprobar si el laboratorio es de fiar se tomó una muestra de 40 dosis, obteniendo una media de 9940 unidades y una desviación estándar de 120 unidades.

   a) ¿Podemos aceptar la indicación del laboratorio con un nivel de significación del 1 %?;
   b) Calcule el nivel de significación de la prueba.

7. En un restaurante se había tenido una media de 160 consumidores diarios y una desviación estándar de 17.5. Se aumentó los precios de la comida y el propietario notó que los 30 últimos días había un promedio de 151 clientes diarios. ¿El propietario puede pensar que efectivamente ha descendido el número de clientes o que la variación es debida al azar?

8. Una empresa exportadora de camarón el año pasado embarcó una media de 10 500 cajas por semana, con una desviación estándar de 1500. En los primeros seis meses (26 semanas) de este año exportó un promedio de 11 200 cajas cada semana.

   a) ¿Puede asegurarse que ha habido un cambio en la demanda de camarón?;
   b) Halle el nivel de significación de la prueba.

9. En una fábrica, la producción por hora de botones de tagua sigue una distribución normal con desviación estándar de 16. Se extrae de dicha población una muestra de tamaño 81, para contrastar la hipótesis de que la producción media por hora del citado producto es 100, al nivel de significación del 5 %. Si la hipótesis alternativa es 105, calcule la probabilidad de cometer un error de tipo II.

10. Una socióloga afirma que el tiempo que los niños de tres a cinco años dedican a ver la televisión cada semana se distribuye normalmente con media 22 horas y desviación estándar 6 horas. Frente a este estudio, una empresa de investigación de mercados cree que la media es mayor y para probar su hipótesis tomó una muestra de 64 observaciones procedentes de la misma población, obteniendo como resultado una media de 25. Si se utiliza un nivel de significación del 5 %:

    a) Verifique si la afirmación del investigador es realmente cierta;
    b) Determine la potencia de ese contraste, si la verdadera media poblacional fuera 24.

11. Un microempresario está considerando la posibilidad de administrar el bar de un colegio. El administrador actual del bar afirma que el ingreso diario sigue una distribución normal de media 87.5 dólares y una desviación estándar de 7.5 dólares. Para comprobar si decía la verdad, tomó una muestra de treinta días y ésta reveló un ingreso diario promedio de 82.5 dólares. Utilizando $\alpha = 0.1$,

    a) ¿hay evidencia de que el ingreso diario promedio sea menor del que afirma el dueño actual?;
    b) Calcule la potencia del contraste si la verdadera media fuera $\mu = 85$.

Pruebas sobre la media poblacional (varianza desconocida)

12. Una muestra de $n = 11$ observaciones de una población distribuida normalmente dio como resultado un promedio $\overline{x} = 19.5$ y una desviación estándar $s = 2.4$. ¿Proporcionan los datos suficiente evidencia para probar que $\mu > 18$?

a) Enuncie las hipótesis nula y alternativa de la prueba;

b) Obtenga la región de rechazo para la prueba si $\alpha = 0.05$;

c) Lleve a cabo la prueba e interprétela.

13. Se desea saber si la edad promedio a la cual se desposan las mujeres en la ciudad de Cuenca es diferente de los 26 años de edad. Se tomó una muestra de 24 registros de matrimonio, resultando las siguientes edades de las novias:

18	28	46	21	29	23	41	43	23	32	20	56
26	40	19	35	20	18	16	65	22	19	24	32

La edad promedio de estas edades es 29.83 años y la desviación estándar es 12.86.

a) ¿Son los 29.83 y 12.86 parámetros o estadísticos?;

b) Plantee las hipótesis nula y alternativa para verificar el argumento arriba indicado;

c) Calcule el estadístico de prueba y la región de rechazo para esta prueba de hipótesis;

d) Explique su conclusión sobre si la edad promedio de las desposadas en la ciudad de Cuenca difiere de los 26 años.

14. Una compañía de televisión por cable anuncia que el periodo medio de espera desde la solicitud hasta la conexión a la red de sus nuevos clientes es de ocho días. Una asociación de consumidores desea verificar dicha hipótesis, para lo cual tomó una muestra del periodo de espera (en días) de 15 clientes con los siguientes resultados:

11 8 10 13 8 10 12 12 8 10 11 7 10 7 8

Verifique si el periodo medio de espera es igual o diferente a ocho días. (Use $\alpha = 0.05$.)

15. La resistencia, en kg/cm^2, de la fibra de carbono se distribuye normalmente. Se tomó una muestra de 10 elementos fabricados con este material, obteniendo:

203 229 215 220 223 233 233 208 228 209

Contraste la hipótesis de que la muestra proviene de una población de media 215 kg/cm^2.

16. Según los datos de una universidad, sus estudiantes obtienían en el examen de evaluación del inglés como segunda lengua una media de 50 puntos. Un profesor de inglés quiso comprobar si sus alumnos tenían un promedio más alto, para ello seleccionó una muestra aleatoria de 20 alumnos y les envió a examinarse. Los resultados dieron una nota promedio de 54 y desviación estándar de 7 puntos. ¿A qué conclusión llegará el profesor, con un nivel de significación del 5 %?

17. En una ciudad se quiere hacer un estudio rápido para valorar el consumo de agua en los domicilios particulares durante los meses de mayor sequía. Para ello se seleccionaron, al azar, 15 domicilios y se midieron sus consumos (x_i) en metros cúbicos durante el mes de agosto. Los resultados fueron $\sum x_i = 280.5$, $\sum x_i^2 = 5308.35$. En vista de estos datos, ¿hay suficiente evidencia estadística, al nivel 0.05, a favor de la hipótesis de que el consumo medio de los particulares durante el mes de agosto es mayor que $18\,\text{m}^3$ (que es el consumo considerado como «sostenible»)?

18. El consumo de gasolina (en litros por 100 km) de los automóviles de 3 puertas sigue una distribución normal con media 8. Se introdujo una modificación en el motor con objeto de disminuir el

consumo y se probaron 10 autos con el motor modificado, obteniéndose los siguientes estadísticos por 100 km:

$$\sum_{i=1}^{10} x_i = 77.5, \quad \sum_{i=1}^{10} x_i^2 = 601.33.$$

Suponiendo que la modificación mantiene la normalidad, ¿hay suficiente evidencia estadística (al nivel de significación 0.05) para poder afirmar que la modificación ha reducido el consumo medio?

19. El tiempo de acceso al disco duro en un cierto modelo de computadoras es una variable aleatoria con media 15 milisegundos. Se propuso una modificación técnica con objeto de disminuir este tiempo de acceso. Se prueba el nuevo sistema en 10 computadoras obteniéndose una media muestral de 14 ms y una desviación estándar de 2.286.

    a) ¿Hay suficiente evidencia estadística, al nivel 0.05, a favor de la hipótesis de que el nuevo modelo disminuye el tiempo de acceso?;
    b) Calcule el $p-$valor.
    c) Indique un par de valores de $\overline{x}$ y de s que hubiesen llevado a un $p-$valor de 0.005. ¿Qué decisión se debería tomar en este caso respecto al nuevo modelo?

20. Una muestra de 25 trabajadores informales, que se dedican a vender golosinas en los buses, ganan un promedio de 17.5 dólares diarios con una desviación estándar de 2.5 dólares.

    a) Estime la ganancia promedio diaria de este tipo de trabajadores informales, usando un intervalo de confianza al 95 %;
    b) ¿Es la ganancia promedio de 17.5 dólares significativamente inferior a los 18.5 dólares que ganan los vendedores que tienen un puesto fijo en los cines? (Utilice un nivel del 5 %.)

Pruebas sobre la varianza poblacional

21. Una muestra de $n = 25$ observaciones de una población distribuida normalmente dio como resultado una varianza muestral de $s^2 = 23.4$. ¿Proporcionan los datos suficiente evidencia que indique que $\sigma^2 > 15$?

22. Una muestra aleatoria de 14 estudiantes de enfermería participan en una prueba para medir su destreza manual, una vez obtenidas las puntuaciones, la varianza de la muestra es de 1225. Contraste con $\alpha = 0.05$ que la varianza poblacional es de 2500.

23. La precisión del trabajo de una máquina automática se mide por la varianza de la dimensión controlada de los productos, que no debe ser mayor que $\sigma_0^2 = 0.18$. Se tomó una muestra aleatoria de artículos que ha dado las siguientes mediciones:

x_i	3.0	3.5	3.8	4.4	4.5
n_i	2	6	9	7	1

Verifique si la máquina garantiza la precisión necesaria para el nivel de significación de 0.05.

24. La desviación estándar de un proceso industrial que produce varillas, en condiciones normales, es de 3 cm. Se dispone de una muestra de tamaño 15, con los siguientes valores:

127 117 118 130 117 122 116 123 121 117 116 122 120 126 115.

Suponiendo que las longitudes siguen una ley normal, contraste la hipótesis de que el proceso funciona correctamente.

25. Una máquina de empaquetado automático deposita en cada paquete una cierta cantidad de papas fritas. Se seleccionan 20 paquetes, se pesa su contenido y se obtienen los siguientes resultados:

49	50	49	50	50	50	49	50	50	50
49	50	50	51	52	48	50	51	51	51

A partir de esta información y suponiendo que la variable se distribuye normalmente:

a) Verifique si la media de esa variable es 51, con un nivel de significación del 1 %;

b) Verifique si la varianza es la unidad, con un nivel de significación del 5 %.

Prueba sobre la proporción poblacional

26. Una muestra de $n = 1000$ observaciones de una población binomial produjo $y = 280$ éxitos. Si la hipótesis de la investigación es que la proporción es menor que 0.3.

a) ¿Es una prueba bilateral o unilateral? ¿Por qué?;

b) Formule las hipótesis nula y alternativa de la prueba;

c) Realice la prueba e interprételа. Utilice $\alpha = 0.05$.

27. Una marca de aceite comestible cubre actualmente el 20 % de los potenciales clientes. Para incrementar las ventas se estructura una campaña publicitaria intensiva. Al final de la misma se realizará una investigación a 400 consumidores potenciales para determinar si ha tenido éxito.

a) Enuncie H_0 y H_1 en términos de p, la probabilidad que el consumidor prefiera la marca de aceite;

b) La empresa concluirá que la campaña tuvo éxito si al menos 94 de los 400 entrevistados prefieren su producto. Encuentre α;

c) Realice la prueba e interprétela.

28. En una encuesta a 300 taxistas, 132 contestaron que utilizan el cinturón de seguridad. Utilizando un nivel de significación del 5 %, ¿podemos concluir que la mitad de los conductores utilizan el cinturón?

29. De acuerdo con sus registros, una clínica ha establecido que la probabilidad de curación completa de un enfermo que ha tomado el medicamento A es 0.8. La clínica experimentó con un nuevo medicamento B en 700 pacientes, de los cuales 575 se curaron totalmente. ¿Se puede considerar que el nuevo medicamento es más eficaz que el tradicional, al nivel de significación de 5 %?

30. Una agencia dedicada al cobro de cheques encontró que el 5 % de todos los cheques remitidos a la agencia eran de cuentas sin fondos. Después de implantar un sistema de verificación, para disminuir sus pérdidas, se hallaron solamente 50 cheques sin fondos en una muestra aleatoria de 1124 cheques. ¿Existe suficiente evidencia estadística para concluir que el sistema de verificación ha reducido la proporción de cheques sin fondos?

31. Un proveedor asegura que los artículos que él suministra son defectuosos en un 1 % de los casos. Se realizó una prueba a 200 de sus artículos y resultaron 3 defectuosos.

a) A un nivel de significación del 1 %, ¿es falsa o cierta la afirmación del proveedor?;

b) Halle el nivel de significación observado de la prueba.

32. En una encuesta en Quito se entrevistó a 850 adultos. A la pregunta de que si ellos estaban a favor del endurecimiento de las penas por ciertos delitos, el 52 % respondió afirmativamente.

a) ¿Se puede concluir que la mayoría de los adultos están a favor del endurecimiento de las penas?;

b) Encuentre el nivel de significación de la prueba.

33. Un economista del Banco Central está interesado en comparar el índice de ausentismo laboral en Quito con el del resto del país, donde se sitúa en el 11 %. Con este propósito seleccionó, al azar, una muestra de 200 trabajadores de dicha ciudad, la cual proporcionó un porcentaje de ausentismo del 16 %.

    a) ¿Se puede sacar la conclusión de que el ausentismo es mayor en Quito que en el resto del país, al nivel de significación $\alpha = 0.025$?;

    b) Calcule β, si $p = 0.17$;

34. Una agencia de publicidad trató de convencer a un industrial para que hiciera propaganda televisada de uno de sus productos, asegurándole que el programa en el que se incluiría su anuncio era visto por el 20 % de las familias. El industrial quiso hacer un experimento por su cuenta. Llamó por teléfono, durante la hora del programa, a 220 familias y halló que en 33 de ellas se veía el programa.

    a) ¿La diferencia entre la proporción de la agencia y la obtenida por el industrial es significativa? Es decir, ¿se debe únicamente al azar o es sustancial? Use $\alpha = 1\,\%$;

    b) Calcule la potencia del contraste, si $p = 0.15$.

35. En un programa periodístico de opinión se pidió que los televidentes llamaran al canal y respondan a la pregunta « ¿cree usted que es necesario que se hagan reformas en el sistema educativo del país?» De 812 llamadas recibidas, 790 se expresaron afirmativamente.

    a) Use esta información para realizar una prueba de que más del 95 % de los ecuatorianos adultos cree que se deben hacer reformas en el sistema educativo;

    b) ¿Es el resultado de la prueba estadísticamente significativa al nivel 0.01?;

    c) Liste las hipótesis requeridas para que los resultados de la prueba sean válidos en esta situación. En este caso, ¿se satisfacen tales hipótesis?

Al igual que en el caso de observaciones realizadas sobre una misma población, cuando se dispone de varias muestras provenientes de poblaciones distintas, se podría desear conocer si ellas comparten alguna característica o si son totalmente diferentes. En lo que sigue nos ocuparemos de las pruebas de hipótesis que involucran dos muestras.

8.7. Pruebas de hipótesis para la diferencia entre dos medias

Igual a lo que se hizo en la sección dedicada a las pruebas de hipótesis sobre la media, se deben considerar los casos en que es posible aplicar el Teorema del Límite Central y aquellos en que no es posible. Adicionalmente, se deberá tener en cuenta un tercer caso, cuando las muestras provienen de una misma unidad muestral, mediante mediciones repetidas.

8.7.1. Caso 1: Varianzas supuestas conocidas

Supongamos que se dispone de dos poblaciones, que las nombraremos como 1 y 2, y se desea probar si la diferencia entre las dos medias poblacionales es igual a una cantidad D_0; es decir, H_0: $\mu_1 - \mu_2 = D_0$, o se desea probar el caso particular de la igualdad de tales medias, o sea, H_0: $\mu_1 = \mu_2$.

De la población 1 se extrae una muestra de tamaño n_1 y de la población 2 se extrae una muestra de tamaño n_2. Si las correspondientes varianzas poblacionales son conocidas, las pruebas de hipótesis son las siguientes:

a) **Prueba bilateral para la diferencia de dos medias.**

1. *Hipótesis Nula.* H_0: $(\mu_1 - \mu_2) = D_0$.
2. *Hipótesis Alternativa.* H_1: $(\mu_1 - \mu_2) \neq D_0$.
3. *Estadístico de Prueba.* $z_{obs} = \dfrac{(\overline{x}_1 - \overline{x}_2) - D_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$.
4. *Región de Rechazo.* $z_{obs} < -z_{\alpha/2}$ o $z_{obs} > z_{\alpha/2}$.

b) **Prueba unilateral para la diferencia de dos medias.**

1. *Hipótesis Nula.* H_0: $(\mu_1 - \mu_2) = D_0$.
2. *Hipótesis Alternativa.* H_1: $(\mu_1 - \mu_2) > D_0$ (o bien H_1: $(\mu_1 - \mu_2) < D_0$).
3. *Estadístico de Prueba.* $z_{obs} = \dfrac{(\overline{x}_1 - \overline{x}_2) - D_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$.
4. *Región de Rechazo.* $z_{obs} > z_\alpha$ (o bien $z_{obs} < -z_\alpha$, cuando H_1: $(\mu_1 - \mu_2) < D_0$).

Ejemplo. Una inversionista tiene dos hoteles en la ciudad, uno en el norte y otro en el sur. Él sospecha que el consumo medio en el restaurante del norte es menor que en el del sur. Del primer local se obtuvo una muestra de 30 facturas, resultando un consumo medio de 59 dólares. Del segundo local se tomó una muestra de 50 facturas, con un consumo medio de 63 dólares. Las varianzas de los consumos en los dos locales son conocidas e iguales a 60 y 80, respectivamente.

a) Para un nivel de significación de 0.05, verifíque si es cierta la sospecha del dueño de los hoteles;

b) Determine la significación de la prueba.

Solución: Por el enunciado del problema:

- Local norte: $\overline{x}_1 = 59$, $\sigma_1^2 = 60$, $n_1 = 30$.
- Local sur: $\overline{x}_2 = 63$, $\sigma_2^2 = 80$, $n_2 = 50$.

a) Es una prueba unilateral con $D_0 = 0$:

1. *Hipótesis Nula.* H_0: $\mu_1 = \mu_2$.
2. *Hipótesis Alternativa.* H_1: $\mu_1 < \mu_2$.

3. *Estadístico de Prueba.* $z_{obs} = \dfrac{(\overline{x}_1 - \overline{x}_2) - D_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} = \dfrac{(59 - 63) - 0}{\sqrt{\dfrac{60}{30} + \dfrac{80}{50}}} = -2.11.$

4. *Región de Rechazo.* Para el nivel de significación $\alpha = 0.05$, la región de rechazo es $z_{obs} < -1.645$.

5. *Decisión.* Como z_{obs}, cae en la región de rechazo; se concluye que en el local del norte el consumo es menor.

b) Puesto que $z_{obs} = -2.11$, entonces $\Phi(-2.11) = 0.0174$; es decir, el $p-$valor es $1.74\,\%$.

Observación. Si el tamaño de las muestras es suficientemente grande ($n_i \geq 25$) y se desconocen las varianzas, se puede utilizar estas pruebas, sustituyendo σ_1^2 y σ_2^2 por sus estimadores, s_1^2 y s_2^2.

8.7.2. Caso 2: Varianzas desconocidas, supuestas iguales

Se dispone de dos poblaciones y se desea probar si la diferencia entre sus correspondientes medias poblacionales es igual a D_0; es decir, H_0: $\mu_1 - \mu_2 = D_0$.

Disponemos de dos muestras de tamaños n_1 y n_2. Si las varianzas poblacionales son desconocidas, pero supuestas iguales; entonces, calculamos el estimador ponderado de σ^2 mediante

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{\sum_{i=1}^{n_1}(x_{1i} - \overline{x}_1)^2 + \sum_{i=1}^{n_2}(x_{2i} - \overline{x}_2)^2}{n_1 + n_2 - 2},$$

donde s_1^2 y s_2^2 son las varianzas muestrales de las muestras 1 y 2.

Con todo esto, las pruebas de hipótesis son las siguientes:

a) **Prueba bilateral para la diferencia de dos medias.**

1. *Hipótesis Nula.* H_0: $(\mu_1 - \mu_2) = D_0$.
2. *Hipótesis Alternativa.* H_1: $(\mu_1 - \mu_2) \neq D_0$.
3. *Estadístico de Prueba.* $t_{obs} = \dfrac{(\overline{x}_1 - \overline{x}_2) - D_0}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$.
4. *Región de Rechazo.* $t_{obs} < -t_{\alpha/2}(n_1 + n_2 - 2)$ o $t_{obs} > t_{\alpha/2}(n_1 + n_2 - 2)$.

b) **Prueba unilateral para la diferencia de dos medias.**

1. *Hipótesis Nula.* H_0: $(\mu_1 - \mu_2) = D_0$.
2. *Hipótesis Alternativa.* H_1: $(\mu_1 - \mu_2) > D_0$ (o bien H_1: $(\mu_1 - \mu_2) < D_0$).
3. *Estadístico de Prueba.* $t_{obs} = \dfrac{(\overline{x}_1 - \overline{x}_2) - D_0}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$.
4. *Región de Rechazo.* $t_{obs} > t_\alpha(n_1 + n_2 - 2)$ (o bien $t_{obs} < -t_\alpha(n_1 + n_2 - 2)$, cuando H_1: $(\mu_1 - \mu_2) < D_0$).

Observación. El supuesto que realizamos –de igualdad entre las varianzas poblaciones–, debe ser comprobado mediante la prueba de hipótesis correspondiente, que la explica en la Sección 8.8.

Ejemplo. Un inversionista no sabe si invertir en bonos emitidos por un país A o por un país B. Para realizar una decisión, seleccionó dos muestras correspondientes a los rendimientos de los bonos emitidos por los dos países, obteniendo los siguientes resultados:

País A

Rendimiento (%)	x_i	12.3	12.5	12.8	13.0	13.5
Frecuencia	n_i	1	2	4	2	1

País B

Rendimiento (%)	y_i	12.2	12.3	13.0
Frecuencia	m_i	6	8	2

A un nivel de significación del 0.01, verificar si el rendimiento de los bonos de los dos países es el mismo. (Asumir que los rendimientos siguen una distribución normal y tienen igual varianza.)

Solución: Se tiene que

$$\overline{x} = 12.80, \quad s_x^2 = 0.11, \quad n_x = 10,$$
$$\overline{y} = 12.35, \quad s_y^2 = 0.07, \quad n_y = 16.$$

El estimador de la varianza es $s^2 = \dfrac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2} = \dfrac{9 \times 0.11 + 15 \times 0.07}{10 + 16 - 2} = \dfrac{1}{12}$.

El contraste es bilateral:

1. *Hipótesis Nula.* H_0: $\mu_x = \mu_y$.
2. *Hipótesis Alternativa.* H_1: $\mu_x \neq \mu_y$.
3. *Estadístico de Prueba.* $t_{obs} = \dfrac{(\overline{x} - \overline{y})}{s\sqrt{\dfrac{1}{n_x} + \dfrac{1}{n_y}}} = \dfrac{12.80 - 12.35}{0.2887\sqrt{\dfrac{1}{10} + \dfrac{1}{16}}} = 3.87.$
4. *Región de Rechazo.* Como $t_{0.005}(24) = 2.797$, la región crítica es $t_{obs} > 2.797$ o $t_{obs} < -2.797$.
5. *Decisión.* Como t_{obs} cae en la región de rechazo, se puede asegurar que los bonos de los dos países tienen rendimientos diferentes. ◄

   Queda como ejercicio para el lector determinar en cuáles bonos se recomienda invertir.

8.7.3. Caso 3: Varianzas desconocidas, supuestas distintas

Supongamos que se dispone de dos poblaciones y se desea probar si la diferencia entre sus correspondientes medias poblacionales es igual a D_0; es decir, H_0: $\mu_1 - \mu_2 = D_0$. Para ello, admitiremos que las poblaciones son normales, cuyas varianzas poblacionales son desconocidas y distintas.

a) **Prueba bilateral para la diferencia de dos medias.**

   1. *Hipótesis Nula.* H_0: $(\mu_1 - \mu_2) = D_0$.

2. *Hipótesis Alternativa.* H_1: $(\mu_1 - \mu_2) \neq D_0$.
3. *Estadístico de Prueba.* $t_{obs} = \dfrac{(\overline{x}_1 - \overline{x}_2) - D_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$.
4. *Región de Rechazo.* $t_{obs} < -t_{\alpha/2}(g)$ o $t_{obs} > t_{\alpha/2}(g)$, donde el número de grados de libertad se calcula por $g = \dfrac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \dfrac{\left(\dfrac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$.
   Cuando g no es un número natural, se redondea al entero más cercano.

b) **Prueba unilateral para la diferencia de dos medias.**

1. *Hipótesis Nula.* H_0: $(\mu_1 - \mu_2) = D_0$.
2. *Hipótesis Alternativa.* H_1: $(\mu_1 - \mu_2) > D_0$ (o bien H_1: $(\mu_1 - \mu_2) < D_0$).
3. *Estadístico de Prueba.* $t_{obs} = \dfrac{(\overline{x}_1 - \overline{x}_2) - D_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$.
4. *Región de Rechazo.* $t_{obs} > t_\alpha(g)$ (o bien $t_{obs} < -t_\alpha(g)$, cuando H_1: $(\mu_1 - \mu_2) < D_0$).

Ejemplo. Se desea conocer el efecto del frío extremo sobre la realización de operaciones manuales. Para ello se eligieron al azar 20 voluntarios, divididos en dos grupos de 10. Al primer grupo se le expuso a una temperatura de 4 °C, mientras que al otro se le mantuvo a temperatura ambiente. Se contabilizó el número de veces que los voluntarios podían abrir y cerrar la mano en un lapso de 15 segundos, con los siguientes resultados:

No expuestos al frío	54	51	40	45	48	46	45	52	49	50
Expuestos al frío	32	29	38	33	34	35	36	35	29	23

Probar la hipótesis que el estar expuesto al frío reduce la capacidad de abrir y cerrar la mano en más de 12 veces.

Solución: Este es un contraste unilateral donde $D_0 = 12$ y

$$\overline{x}_1 = 48.0, \quad s_1^2 = 16.89, \quad n_1 = 10,$$
$$\overline{x}_2 = 32.4, \quad s_2^2 = 19.16, \quad n_2 = 10.$$

1. *Hipótesis Nula.* H_0: $\mu_1 - \mu_2 = 12$.
2. *Hipótesis Alternativa.* H_1: $\mu_1 - \mu_2 > 12$.
3. *Estadístico de Prueba.* $t_{obs} = \dfrac{(\overline{x}_1 - \overline{x}_2) - D_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \dfrac{(48 - 32.4) - 12}{\sqrt{\dfrac{16.89}{10} + \dfrac{19.16}{10}}} = 1.896\,1.$
4. *Región de Rechazo.* El número de grados de libertad es

$$g = \frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \dfrac{\left(\dfrac{s_2^2}{n_2}\right)^2}{n_2 - 1}} = \frac{\left(\dfrac{16.89}{10} + \dfrac{19.16}{10}\right)^2}{\dfrac{\left(\dfrac{16.89}{10}\right)^2}{10 - 1} + \dfrac{\left(\dfrac{19.16}{10}\right)^2}{10 - 1}} = 17.9 \approx 18.$$

Entonces, $t_{0.05}(18) = 1.734$ y la región crítica es $t_{obs} > 1.734$.

5. *Decisión.* Como t_{obs} cae en la región de rechazo del contraste se puede admitir que el frío reduce la capacidad de realizar operaciones manuales.

8.7.4. Caso 4: Una varianza conocida y otra varianza desconocida

Supongamos que se dispone de dos poblaciones y se desea probar si la diferencia entre sus medias poblacionales es igual a D_0; es decir, H_0: $\mu_1 - \mu_2 = D_0$. Admitiremos que las poblaciones son normales y, sin pérdida de generalidad, supondremos que la varianza σ_1^2 es conocida y que la varianza σ_2^2 es desconocida.

a) **Prueba bilateral para la diferencia de dos medias.**

1. *Hipótesis Nula.* H_0: $(\mu_1 - \mu_2) = D_0$.
2. *Hipótesis Alternativa.* H_1: $(\mu_1 - \mu_2) \neq D_0$.
3. *Estadístico de Prueba.* $t_{obs} = \dfrac{(\overline{x}_1 - \overline{x}_2) - D_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$.
4. *Región de Rechazo.* $t_{obs} < -t_{\alpha/2}(g)$ o $t_{obs} > t_{\alpha/2}(g)$, donde el número de grados de libertad se calcula por $g = \dfrac{\left(\dfrac{\sigma_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$.

   Cuando g no es un número natural, se redondea al entero más cercano.

b) **Prueba unilateral para la diferencia de dos medias.**

1. *Hipótesis Nula.* H_0: $(\mu_1 - \mu_2) = D_0$.
2. *Hipótesis Alternativa.* H_1: $(\mu_1 - \mu_2) > D_0$ (o bien H_1: $(\mu_1 - \mu_2) < D_0$).
3. *Estadístico de Prueba.* $t_{obs} = \dfrac{(\overline{x}_1 - \overline{x}_2) - D_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$.
4. *Región de Rechazo.* $t_{obs} > t_\alpha(g)$ (o bien $t_{obs} < -t_\alpha(g)$, cuando H_1: $(\mu_1 - \mu_2) < D_0$).

8.7.5. Caso 5: Diferencia por parejas

Las pruebas para las diferencias de las medias, realizadas anteriormente, se aplican cuando las dos muestras son independientes, pero existen casos en los que la información recogida no es independiente (como cuando se la toma de un mismo individuo de manera repetida).

Sea (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) una muestra aleatoria de pares de observaciones; donde (x_i, y_i) representa dos mediciones tomadas de la misma unidad muestral, antes y después de un tratamiento o fenómeno que la afectó. Se desea conocer si la población cambió de manera apreciable después del fenómeno indicado; para ello se emplea la prueba de diferencias por parejas de la manera que a continuación se describe.

Se construye una muestra aleatoria de las diferencias $d_1, d_2, \ldots, d_n$, donde $d_i = x_i - y_i$ $(i = 1, 2, \ldots, n)$, que las supondremos siguen una ley normal de media μ_d y varianza σ_d^2. Para estos parámetros poblacionales se calculan sus estimadores:

$$\overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \quad \text{y} \quad s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \overline{d})^2.$$

Prueba bilateral para la diferencia por parejas.

1. *Hipótesis Nula.* H_0: $\mu_d = D_0$.
2. *Hipótesis Alternativa.* H_1: $\mu_d \neq D_0$.
3. *Estadístico de Prueba.* $t_{obs} = \dfrac{\overline{d} - D_0}{s_d/\sqrt{n}}$.
4. *Región de Rechazo.* $t_{obs} < -t_{\alpha/2}(n-1)$ o $t_{obs} > t_{\alpha/2}(n-1)$.

Observación. También, se pueden realizar los contrastes unilaterales considerando las hipótesis alternativas $\mu_d < D_0$ o $\mu_d > D_0$; para cada caso se escogerá, de la manera antes indicada, la correspondiente región de rechazo. Se recomienda que el lector formule tales pruebas.

Ejemplo. En una investigación clínica se realizaron mediciones de la frecuencia cardiaca a 9 individuos antes y después de que ellos se hubieran sometido a un programa de entrenamiento físico. Los resultados son los siguientes:

Antes (x_i)	81	85	82	82	90	83	73	93	92
Después (y_i)	76	71	87	79	90	89	76	75	89

Para un nivel de significación de 0.05, establecer si el acondicionamiento físico varió de manera significativa la frecuencia cardiaca, suponiendo una distribución normal de las diferencias.

Solución: El promedio y la desviación estándar de las diferencias son:

$$\overline{d} = \frac{29}{9} = 3.22, \qquad s_d = 8.21.$$

Además, $D_0 = 0$ y la prueba bilateral queda como sigue:

1. *Hipótesis Nula.* H_0: $\mu_d = 0$.
2. *Hipótesis Alternativa.* H_1: $\mu_d \neq 0$.
3. *Estadístico de Prueba.* $t_{obs} = \dfrac{\overline{d}}{s_d/\sqrt{n}} = \dfrac{3.22}{8.21/\sqrt{9}} = 1.177$.
4. *Región de Rechazo.* Como $t_{0.025}(8) = 2.306$, entonces la región es $t_{obs} > 2.306$ o $t_{obs} < -2.306$.
5. *Decisión.* Como t_{obs} no está en la región crítica, no se rechaza H_0; es decir, no hay por qué considerar que hubo una variación apreciable en la frecuencia cardiaca.

8.8. Pruebas de hipótesis para la razón entre dos varianzas

Supongamos que se desea probar la igualdad de las varianzas de dos poblaciones normalmente distribuidas, de las que se han extraído dos muestras independientes; es decir, se desea probar H_0: $\sigma_1^2 = \sigma_2^2$.

Las pruebas de hipótesis son las siguientes:

a) **Prueba bilateral para la razón entre dos varianzas.**

1. *Hipótesis Nula.* H_0: $\sigma_1^2 = \sigma_2^2$.
2. *Hipótesis Alternativa.* H_1: $\sigma_1^2 \neq \sigma_2^2$.
3. *Estadístico de Prueba.* $F_{obs} = \dfrac{s_1^2}{s_2^2}$, donde s_1^2 es la mayor varianza muestral.
4. *Región de Rechazo.* $F_{obs} > F_{\alpha/2}(n_1 - 1, n_2 - 1)$.

b) **Prueba unilateral para la razón entre dos varianzas.**

1. *Hipótesis Nula.* H_0: $\sigma_1^2 = \sigma_2^2$.
2. *Hipótesis Alternativa.* H_1: $\sigma_1^2 > \sigma_2^2$.
3. *Estadístico de Prueba.* $F_{obs} = \dfrac{s_1^2}{s_2^2}$, donde s_1^2 es la mayor varianza muestral.
4. *Región de Rechazo.* $F_{obs} > F_{\alpha}(n_1 - 1, n_2 - 1)$

Ejemplo (Continuación). Un inversionista no sabe si invertir en bonos emitidos por un país A o por un país B. Para realizar una decisión, seleccionó dos muestras, correspondientes a los rendimientos de los bonos emitidos por los dos países, obteniendo unas varianzas muestrales iguales a $s_x^2 = 0.11$ y $s_y^2 = 0.07$. A un nivel de significación del 0.05, verificar si las varianzas poblacionales de los rendimientos de los bonos de los dos países son iguales.

Solución: La prueba es bilateral:

1. *Hipótesis Nula.* H_0: $\sigma_x^2 = \sigma_y^2$.
2. *Hipótesis Alternativa.* H_1: $\sigma_x^2 \neq \sigma_y^2$.
3. *Estadístico de Prueba.* $F_{obs} = \dfrac{s_x^2}{s_y^2} = \dfrac{0.11}{0.07} = 1.571$.
4. *Región de Rechazo.* Según el nivel de significación 0.05, y los grados de libertad $n_1 - 1 = 10 - 1 = 9$ y $n_2 - 1 = 16 - 1 = 15$, el punto crítico es $F_{0.025}(9, 15) = 3.12$ y define la región $F_{obs} > 3.12$.
5. *Decisión.* Puesto que $1.571 < 3.12$, no se debe rechazar la hipótesis de la igualdad de las varianzas. Entonces, fue correcto asumir que las varianzas eran iguales, cuando realizamos la prueba sobre la igualdad de las medias.

8.9. Pruebas de hipótesis para la diferencia entre dos proporciones

Supongamos que se han seleccionado dos muestras, de manera aleatoria e independiente, de dos poblaciones binomiales, cuyos tamaños, n_1 y n_2 son suficientemente altos para que las distribuciones muestrales de $\widehat{p}_1$ y $\widehat{p}_2$ sean aproximadamente normales. Se desea probar si la diferencia de las proporciones muestrales es igual a un valor D_0. Se deben tomar en cuenta dos casos: cuando $D_0 = 0$ (igualdad de las proporciones) y cuando $D_0 \neq 0$.

8.9.1. Prueba para $p_1 - p_2$ cuando $D_0 = 0$

En primer lugar se estiman las proporciones p_1 y p_2, a partir de las muestras, por $\widehat{p}_1$ y $\widehat{p}_2$, respectivamente, luego se calcula el *estimador ponderado de la proporción* p por

$$\widehat{p} = \frac{n_1\widehat{p}_1 + n_2\widehat{p}_2}{n_1 + n_2}.$$

Prueba bilateral para la diferencia entre dos proporciones.

1. *Hipótesis Nula.* H_0: $(p_1 - p_2) = 0$, es decir H_0: $p_1 = p_2 = p$.
2. *Hipótesis Alternativa.* H_1: $(p_1 - p_2) \neq 0$.
3. *Estadístico de Prueba.* $z_{obs} = \dfrac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}\widehat{q}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$.
4. *Región de Rechazo.* $z_{obs} < -z_{\alpha/2}$ o $z_{obs} > z_{\alpha/2}$.

Observación. También, se pueden realizar las pruebas unilaterales para probar que H_1: $p_1 > p_2$ o que H_1: $p_1 < p_2$, cambiando las regiones de rechazo según cada caso particular.

Ejemplo. El dueño de un supermercado cree que el porcentaje de cheques protestados, con que los clientes han pagado sus cuentas, ha aumentado con respecto al año anterior. En una muestra correspondiente al primer trimestre del año pasado, encontró 5 cheques protestados de 80 cheques admitidos; con otra muestra de 68 cheques, correspondientes al primer trimestre del presente año, el número de cheques protestados fue de 6. Con estos datos, ¿hay evidencia suficiente que indique un incremento en el porcentaje de cheques protestados?

Solución: Sean p_1 y p_2 las proporciones de cheques protestados en el primer trimestre del año anterior y del presente. Estimemos las proporciones muestrales:

$$\widehat{p}_1 = \frac{5}{80} = 0.0625, \qquad \widehat{p}_2 = \frac{6}{68} = 0.0882,$$

$$\widehat{p} = \frac{5+6}{80+68} = 0.0743.$$

Ya que se desea determinar si existe un aumento en la proporción, la hipótesis alternativa es H_1: $p_2 - p_1 > 0$, y la prueba es la siguiente:

1. *Hipótesis Nula.* H_0: $p_1 = p_2$.
2. *Hipótesis Alternativa.* H_1: $p_1 < p_2$.
3. *Estadístico de Prueba.* $z_{obs} = \dfrac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}\widehat{q}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} = \dfrac{0.0625 - 0.0882}{\sqrt{(0.0743)(0.9257)\left(\dfrac{1}{80} + \dfrac{1}{68}\right)}} = -0.594.$
4. *Región de Rechazo.* Tomemos $\alpha = 0.05$; entonces, la región es $z_{obs} < -1.645$.
5. *Decisión.* Como $-0.594 > -1.645$, no se puede rechazar la hipótesis nula; o sea, no hay evidencia que indique un aumento en el número de cheques protestados, con respecto al año anterior.

8.9.2. Prueba para $p_1 - p_2$ cuando $D_0 \neq 0$

En el caso en que se desee probar que las dos proporciones son distintas y que su diferencia es igual a un valor dado, D_0, se tiene la siguiente prueba:

Prueba bilateral para la diferencia entre dos proporciones.

1. *Hipótesis Nula.* H_0: $(p_1 - p_2) = D_0$.
2. *Hipótesis Alternativa.* H_1: $(p_1 - p_2) \neq D_0$.
3. *Estadístico de Prueba.* $z_{obs} = \dfrac{(\widehat{p}_1 - \widehat{p}_2) - D_0}{\sqrt{\dfrac{\widehat{p}_1\widehat{q}_1}{n_1} + \dfrac{\widehat{p}_2\widehat{q}_2}{n_2}}}$.
4. *Región de Rechazo.* $z_{obs} < -z_{\alpha/2}$ o $z_{obs} > z_{\alpha/2}$.

Observación. También, se pueden realizar las pruebas unilaterales para probar que H_1: $(p_1 - p_2) > D_0$ o que H_1: $(p_1 - p_2) < D_0$, cambiando las regiones de rechazo según cada caso particular.

Ejemplo. En el deporte del balonmano, en un partido está permitido sustituir al portero solo para que detenega los tiros penalties. El entrenador de un equipo, al definir su estrategia para un partido, examina las estadísticas individuales de los porteros titular y suplente. En una muestra de los registros de los entrenamientos del último mes, el titular ha detenido 128 de 510 penalties y el suplente ha detenido 183 de 480 tiros. El entrenador decidirá sustituir al portero titular, el momento de parar un penalti, si el suplente ha detenido al menos un 10 % más de tiros que el titular. A un nivel de significación del 5 %, ¿qué decisión tomará el entrenador?

Solución: Calculemos las proporciones correspondientes a cada uno de los porteros:

$$\widehat{p}_1 = \frac{128}{510} = 0.251, \qquad \widehat{p}_2 = \frac{183}{480} = 0.381.$$

Se realizará la prueba unilateral con $D_0 = 0.1$.

1. *Hipótesis Nula.* H_0: $p_2 - p_1 = 0.1$.
2. *Hipótesis Alternativa.* H_1: $p_2 - p_1 > 0.1$.
3. *Estadístico de Prueba.* $z_{obs} = \dfrac{(\widehat{p}_2 - \widehat{p}_1) - D_0}{\sqrt{\dfrac{\widehat{p}_1\widehat{q}_1}{n_1} + \dfrac{\widehat{p}_2\widehat{q}_2}{n_2}}} = \dfrac{(0.381 - 0.251) - 0.1}{\sqrt{\dfrac{0.251 \times 0.749}{510} + \dfrac{0.381 \times 0.619}{480}}} = 1.023.$
4. *Región de Rechazo.* Al nivel de significación 0.05, la región de rechazo es $z_{obs} > 1.645$.
5. *Decisión.* Como $z_{obs} = 1.023$, no cae en la región de rechazo, no hay razón para pensar que la diferencia es mayor que 0.1; entonces, el entrenador no debería decidirse a sustituir al portero titular.

En el Cuadro 8.2 se presenta un resumen de las pruebas de hipótesis con dos muestras.

Población	Hipótesis nula (H_0)	Hipótesis alternativa (H_1)	Estadístico de prueba	Región de rechazo
General con varianzas iguales	$\mu_1 - \mu_2 = D_0$ $\mu_1 - \mu_2 \leq D_0$ $\mu_1 - \mu_2 \geq D_0$	$\mu_1 - \mu_2 \neq D_0$ $\mu_1 - \mu_2 > D_0$ $\mu_1 - \mu_2 < D_0$	$z = \dfrac{(\overline{x}_1 - \overline{x}_2) - D_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$	$\|z\| > z_{\alpha/2}$ $z > z_\alpha$ $z < -z_\alpha$
Normal varianzas desconocidas supuestas iguales	$\mu_1 - \mu_2 = D_0$ $\mu_1 - \mu_2 \leq D_0$ $\mu_1 - \mu_2 \geq D_0$	$\mu_1 - \mu_2 \neq D_0$ $\mu_1 - \mu_2 > D_0$ $\mu_1 - \mu_2 < D_0$	$t = \dfrac{(\overline{x}_1 - \overline{x}_2) - D_0}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ $n_1 + n_2 - 2$ g.l.	$\|t\| > t_{\alpha/2}$ $t > t_\alpha$ $t < -t_\alpha$
Normal varianzas desconocidas supuestas distintas	$\mu_1 - \mu_2 = D_0$ $\mu_1 - \mu_2 \leq D_0$ $\mu_1 - \mu_2 \geq D_0$	$\mu_1 - \mu_2 \neq D_0$ $\mu_1 - \mu_2 > D_0$ $\mu_1 - \mu_2 < D_0$	$t = \dfrac{(\overline{x}_1 - \overline{x}_2) - D_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$ r g.l.	$\|t\| > t_{\alpha/2}$ $t > t_\alpha$ $t < -t_\alpha$
Normal observaciones emparejadas	$\mu_D = \mu_{D_0}$ $\mu_D \leq \mu_{D_0}$ $\mu_D \geq \mu_{D_0}$	$\mu_D \neq \mu_{D_0}$ $\mu_D > \mu_{D_0}$ $\mu_D < \mu_{D_0}$	$t = \dfrac{\overline{D} - \mu_{D_0}}{s_d/\sqrt{n}}$ $n - 1$ g.l.	$\|t\| > t_{\alpha/2}$ $t > t_\alpha$ $t < -t_\alpha$
Normal	$\sigma_1^2 = \sigma_2^2$ $\sigma_1^2 \leq \sigma_2^2$ $\sigma_1^2 \geq \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$ $\sigma_1^2 > \sigma_2^2$ $\sigma_1^2 < \sigma_2^2$	$F = \dfrac{s_1^2}{s_2^2}$ $\nu_1 = n_1 - 1, \nu_2 = n_2 - 1$	$F > F_{\alpha/2}$ $F > F_\alpha$ $F < F_{1-\alpha}$
Binomial	$p_1 = p_2$ $p_1 \leq p_2$ $p_1 \geq p_2$	$p_1 \neq p_2$ $p_1 > p_2$ $p_1 < p_2$	$z = \dfrac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}\widehat{q}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ $\widehat{p} = \dfrac{n_1\widehat{p}_1 + n_2\widehat{p}_2}{n_1 + n_2}$	$\|z\| > z_{\alpha/2}$ $z > z_\alpha$ $z < -z_\alpha$
Binomial ($D_0 \neq 0$)	$p_1 - p_2 = D_0$ $p_1 - p_2 \leq D_0$ $p_1 - p_2 \geq D_0$	$p_1 - p_2 \neq D_0$ $p_1 - p_2 > D_0$ $p_1 - p_2 < D_0$	$z = \dfrac{(\widehat{p}_1 - \widehat{p}_2) - D_0}{\sqrt{\dfrac{\widehat{p}_1\widehat{q}_1}{n_1} + \dfrac{\widehat{p}_2\widehat{q}_2}{n_2}}}$	$\|z\| > z_{\alpha/2}$ $z > z_\alpha$ $z < -z_\alpha$

Cuadro 8.2: Pruebas de hipótesis con dos muestras

8.10. Ejercicios

Pruebas sobre la diferencia de dos medias (varianzas conocidas)

1. Dos máquinas envasan cereal en cajas. De la primera máquina se obtuvo una muestra de 30 cajas, resultando un peso promedio de 130 g y de la segunda máquina se tomó una muestra de 50 cajas, con un peso promedio de 125 g. Las varianzas de los pesos envasados por las dos máquinas son conocidas e iguales a 60 y 80, respectivamente. Para un nivel de significación de 0.05, verifique si las dos máquinas envasan iguales cantidades de cereal.

2. En una investigación del Ministerio de Salud se midió el contenido de nicotina de dos marcas de cigarrillo. En un experimento con 50 cigarrillos de la primera marca se encontró que tiene un contenido promedio de 2.47 mg con desviación estándar de 0.12 mg; mientras que para 40 cigarrillos de la segunda marca el contenido promedio fue de 2.39 mg con desviación estándar de 0.14 mg. Pruebe la hipótesis nula $\mu_1 - \mu_2 = 0.12$ contra la alternativa $\mu_1 - \mu_2 \neq 0.12$, usando $\alpha = 0.05$.

3. En una ciudad operan 2 empresas de telefonía celular: Sirius y Quark. Por los registros las empresas saben que el gasto mensual de sus abonados tienen desviaciones estándar de 6.5 y 5 dólares, respectivamente. Para comparar el gasto medio de los clientes de las dos empresas, se tomó una muestra, al azar, de 34 clientes de Sirius y otra muestra de 41 clientes de Quark. Los correspondientes consumos mensuales fueron 35 y 38.5 dólares.

a) ¿Proporcionan estos datos evidencia estadística, al nivel 0.04, a favor de la hipótesis de que el gasto de los clientes de Sirius es menor que es gasto de los clientes de Quark?;

b) Halle el $p-$valor de la prueba.

4. Los ingresos del primer empleo de los ingenieros informáticos, egresados de cualquier universidad, siguen una distribución normal con desviación estándar de 3800 dólares. Se tomó una muestra aleatoria de 15 ingenieros procedentes de la Universidad Nacional, resultando que en su primer empleo los ingresos medios anuales fueron de 12 000 dólares. Otra muestra independiente de 12 ingenieros de la Universidad Técnica dio como resultado unos ingresos medios en el primer empleo de 13 200 dólares. Se pide, con un nivel de significación del 2 %, probar la hipótesis de que las medias son iguales frente a la alternativa de que la media de la Universidad Nacional es menor que la de la Universidad Técnica.

5. Una cooperativa agrícola produce cierto arroz con fertilizante natural y con abono químico. En las parcelas donde se emplea fertilizante natural se obtienen plantas cuya altura tiene varianza de $47\,\text{cm}^2$. En los terrenos donde se usa abono químico la altura de las plantas tiene un varianza igual a $39\,\text{cm}^2$. Para comprobar las medias se toma aleatoriamente una muestra de 65 plantas, 31 correspondientes al primer tipo de tierras y 34 al segundo; obteniéndose en las muestras 92 cm y 86 cm de alturas medias, respectivamente. Para un nivel de significación del 6 %, contraste la hipótesis de que los fertilizantes son igualmente eficaces, frente a la hipótesis alternativa de que es más eficaz el natural.

6. Se quiere comparar el rendimiento académico en matemática de los alumnos de último año de dos colegios; A y B. Elegidas dos muestras aleatorias de cada una de estos colegios (30 alumnos de A y 36 de B), se obtuvieron los siguientes resultados en cuanto a la calificación obtenida en dicha asignatura:

	Promedio	**Varianza**
Colegio A	13.8	1.1
Colegio B	14.3	1.6

Basándose en estos resultados, verifique la hipótesis de igualdad de rendimientos académicos de ambos colegios, con un nivel de significación del 3 %.

7. En un proceso químico para producir oxígeno (O_2) se emplea un catalizador. Durante treinta días se midió la cantidad promedio de oxígeno producido luego de haber colocado el catalizador y cuatro horas después, resultando que, en el primer caso se producían 1000 litros de O_2 en una hora, con una desviación estándar de 90 litros y en el segundo caso se producían 880 litros de O_2 con una desviación estándar de 140 litros. ¿Existe evidencia que indique que el catalizador se degrada, produciendo una merma de al menos 100 litros de O_2, cuando han transcurrido cuatro horas?

8. Se efectuó un análisis sobre la duración de las máquinas computadoras que se utilizan en las empresas públicas. Se eligieron dos muestras de computadoras, de marca y de clones, cada una constituida por 80 máquinas. Para las de marca, resultó una vida promedio de 4.8 años y una desviación estándar de 1.7 años. Para las clones, dio una vida promedio de 3.3 años y desviación estándar de 1.2 años. ¿Puede considerarse que la vida media de las computadoras de marca es superior en al menos un año a las que son clones?

9. La FIFA realizó un cambio en la forma de puntuación en los partidos de fútbol ganados: se otorgá tres puntos al equipo ganador, en lugar de los dos puntos usuales. Para examinar la efectividad de la nueva norma, se examinó los resultados de los partidos jugados en 1995 y 1996 (año en el que entró en vigencia la nueva norma). En 45 partidos examinados, jugados el año 1995, se encontró un promedio de 2.87 goles por partido y una desviación estándar de 0.21; en los 38 partidos examinados de 1996, se halló que se habían producido un promedio de 3.05 goles con

desviación estándar de 0.18. ¿Puede decirse que la nueva norma permitió aumentar el promedio de goles por partido?

10. Una persona desea comprar un automóvil nuevo y toma como factor de decisión el consumo medio de combustible que tienen dos modelos de características similares, uno fabricado en Europa y otro en Corea. Consulta una revista especializada y encuentra que el auto europeo tiene un recorrido promedio de 28.3 km por galón de combustible, con una desviación estándar de 6.2 km; para el auto coreano encuentra que el recorrido es de 26.7 km por galón y una desviación estándar de 5.1 km; además, la revista indica que los datos estadísticos fueron tomados a partir de las mediciones realizadas en 50 autos de cada origen. ¿Cuál de los dos autos deberá comprar? o ¿deberá tomar en consideración otras características, distintas del rendimiento del combustible, para tomar su decisión?

Pruebas sobre la diferencia de dos medias (varianzas desconocidas)

11. Se realizaron pruebas para conocer la cantidad de plomo en la sangre de personas expuestas a la contaminación en la ciudad. Se tomaron muestras aleatorias de 24 niños y 18 adultos que dieron los siguientes resultados (en ppm):

Adultos	Niños
$\overline{x}_1 = 0.043$,	$\overline{x}_2 = 0.028$,
$s_1 = 0.018$,	$s_2 = 0.007$.

Asumiendo que las varianzas poblacionales son iguales, pruebe la hipótesis de que no hay diferencia entre el contenido medio de plomo en la sangre de los niños y de los adultos, contra la hipótesis de que los adultos tienen mayor cantidad que los niños.

12. En un estudio sobre un nuevo programa piloto para el aprendizaje a distancia mediante computadora, se eligieron al azar, 21 estudiantes de una clase para seguir el nuevo programa (grupo piloto) mientras que los 23 restantes seguían el método tradicional (grupo control). Finalizado el curso, se realizó un examen obteniéndose los siguientes resultados:

	Grupo piloto	Grupo control
Nota promedio	51.48	41.52
Desv. estándar	11.01	14.15

Suponiendo igualdad de varianzas, contraste si hay evidencia (a nivel del 5 %) de que el nuevo método piloto da mejores resultados que el método tradicional.

13. En el Departamento de Genética de una universidad se realizó un estudio donde se contó el número de surcos en las huellas dactilares de dos grupos de voluntarios: uno de aborígenes amazónicos y otro de estudiantes mestizos de la universidad. Los promedios y la desviación estándar del total de surcos contados se dan a continuación:

$$\begin{array}{llll} \text{aborígenes:} & n_x = 17, & \overline{x} = 125.9, & s_x^2 = (50.20)^2, \\ \text{mestizos:} & n_y = 17, & \overline{y} = 154.3, & s_y^2 = (29.04)^2. \end{array}$$

Pruebe la hipótesis de que no existe diferencia entre el total de surcos, contra una hipótesis adecuada a los datos, asumiendo igualdad en las varianzas.

14. Se probaron 2 medicamentos A y B para la eliminación de los hongos en la piel. En pruebas de laboratorio se tomaron muestras de la cantidad de medicamento requerido para la eliminación de un mismo tipo de hongo, obteniendo los siguientes resultados:

	Medicamento A	Medicamento B
Tamaño	13	15
Promedio (g)	4.5	3.1
Desv. estándar (g)	1.8	1.7

a) Pruebe la hipótesis de que los dos medicamentos tienen igual efecto, contra la hipótesis de que el medicamento B es más efectivo. ¿Qué conclusión saca?;

b) Encuentre el nivel de significación aproximado de la prueba.

15. Dos empresas competidoras (S y T) que venden implementos deportivos han puesto en marcha, casi simultáneamente, páginas de internet para la venta electrónica. Se eligieron, al azar, ocho clientes que visitaron la página S y, de manera independiente, otros ocho que visitaron la T y se midió el tiempo (en minutos) de la duración de la visita de cada cliente. Los resultados fueron los siguientes:

Página S	2.3	3.5	4.2	3.2	4.4	2.1	1.6	5.3
Página T	1.3	2.3	4.4	3.7	2.8	6.5	3.6	4.5

¿Proporcionan estos datos suficiente evidencia (al nivel 0.05) para afirmar que los tiempos medios de duración de las visitas en ambas páginas son diferentes?

16. Para los tres primeros meses del año, 15 vendedores de la Costa tuvieron ventas semanales promedio de 300 dólares con una desviación estándar de 50 dólares; en tanto, 10 vendedores de la Sierra tuvieron ventas semanales promedio de 260 dólares, con una desviación estándar de 16 dólares. Si consideramos que las desviaciones estándar de las ventas son diferentes, determine si los vendedores de la Costa tienen mayores ventas semanales.

17. Se compararon los tiempos (en segundos) que realizan atletas escolares, masculinos y femeninos, al recorrer una distancia de 500 metros. Para el efecto, se registraron las marcas de 9 niños y 7 niñas, obteniendo lo siguiente:

Niños	187	223	235	192	259	175	206	194	247
Niñas	248	366	223	326	274	369	301		

Determine si la diferencia en el tiempo medio entre hombres y mujeres es mayor que un minuto. Use $\alpha = 10\,\%$.

18. Se sospecha que la concentración media de dióxido de carbono es mayor en la capa de aire más próxima a la superficie. Para contrastar esta hipótesis se analizó el aire en 20 puntos elegidos aleatoriamente a un metro de altura del suelo, resultando una media muestral de 580 p.p.m.v. (partes por millón en volumen) y una desviación estándar de 60. También, se realizaron 16 mediciones de la concentración, a una altura de 18 metros, resultando una concentración de 365 p.p.m.v. y una desviación estándar de 110. Suponiendo normalidad para las mediciones y que las varianzas son diferentes, ¿proporcionan estos datos suficiente evidencia, al nivel 0.01, a favor de la hipótesis de que la concentración es mayor cerca del suelo?

Pruebas sobre la diferencia por parejas

19. Se hizo un estudio para comparar los tiempos de acceso, en diferentes momentos del día, a internet desde computadoras domésticas con módem. Para ello, se cargaron 8 páginas web por la tarde en el periodo de 14 a 15 h. y, con la misma máquina, las mismas 8 páginas por la noche en el periodo de 22 a 23 h. Los respectivos tiempos de acceso en minutos fueron:

De 14 a 15 h	2.3	1.5	1.0	2.7	1.4	1.9	0.8	1.1
De 22 a 23 h	2.9	1.4	1.2	3.4	1.3	2.5	1.6	1.8

¿Se puede concluir, al nivel 0.01, que el acceso es más lento en el horario nocturno?

20. Para poner a prueba un nuevo método de estudio, se seleccionó a 10 sujetos que fueron sometidos a una prueba antes y después de entrenarse con el nuevo método. Las calificaciones fueron:

Antes	19	14	23	25	18	24	17	19	20	11
Después	20	15	26	24	17	26	18	22	26	11

A un nivel de significación del 5 %, ¿se puede decir que el nuevo método es efectivo?

21. Un grupo de investigadores afirma haber descubierto un tipo de alimentación para las gallinas, bajo la cual éstas producen huevos que no aumentan el colesterol en las personas que los consumen. Para comprobar dicha teoría, se seleccionaron al azar 36 personas a las que se les midió su nivel de colesterol habitual (x_i). Después someter a estas mismas personas a una dieta a base de los huevos en estudio, se midió en ellas de nuevo dicho nivel (y_i). Los resultados fueron:

$$\overline{x} = 203, \quad \overline{y} = 200, \quad \frac{1}{35}\sum_{i=1}^{36}\left(d_i - \overline{d}\right)^2 = 196.$$

Suponiendo normalidad en la variable, contraste la hipótesis de que los huevos modifican el colesterol, a un nivel de significación del 1 %.

22. Se quiere comparar dos métodos rápidos para estimar la carga eléctrica en un circuito. En un laboratorio se midieron 8 circuitos, mediante los dos métodos, obteniendo los siguientes resultados:

Circuito	1	2	3	4	5	6	7	8
Método A	10.7	11.2	15.3	14.9	13.9	15.0	15.6	15.7
Método B	11.1	11.4	15.0	15.1	14.3	15.4	15.4	16.0

Contraste si los dos métodos proporcionan, en media, las mismas estimaciones.

23. En el artículo «β−Endorphin: A Factor in 'Fun Run' Collapse» (British Medical Journal (1987), 294, 1002), se reportan las siguientes mediciones de la concentración en la sangre de β−endorfina, en picomoles por litro, realizadas a 11 corredores antes y después de haber participado en una media maratón:

Corredor	1	2	3	4	5	6	7	8	9	10	11
Antes de la carrera	4.3	4.6	5.2	5.2	6.6	7.2	8.4	9.0	10.4	14.0	17.8
Después de la carrera	29.6	25.1	15.5	29.6	24.1	37.8	20.2	21.9	14.2	34.6	46.2

Si se sospecha que luego de realizar un esfuerzo prolongado, la concentración de la hormona se eleva en 25 picomoles. Verifique, a un nivel de significación del 5 %, si se cumple la hipótesis.

Pruebas sobre la razón de dos varianzas

24. Se seleccionaron dos muestras aleatorias provenientes de poblaciones normales:

Población	Tamaño de la muestra	Varianza muestral
X	16	56.7
Y	20	30.5

a) ¿Hay suficiente evidencia para pensar que $\sigma_X^2 = \sigma_Y^2$?, a un nivel de significación de 0.05;

b) Encuentre el nivel de significación de la prueba, e interprete el resultado.

25. Una fábrica de refrescos tiene dos máquinas embotelladoras que envasan el líquido en las botellas. Los expendedores han informado al fabricante que las botellas presentan demasiada variabilidad en la cantidad de líquido que contienen. Para examinar la queja se midió la varianza del volumen de líquido embotellado por cada una de las máquinas, resultando que con una muestra de 18 botellas de la máquina 1, se obtuvo una varianza de 700 y con una muestra de 15 botellas de la máquina 2, se obtuvo una varianza de 210. ¿La variabilidad en los volúmenes envasados puede ser atribuida a una o a las dos máquinas? Utilice $\alpha = 0.05$.

26. Dos estaciones meteorológicas predicen la temperatura a medio día en la ciudad con seis horas de anticipación. Se tomaron dos muestras independientes de los datos de cada estación, resultando:

$$\begin{array}{llll} \textbf{Estación 1} & n_1 = 8, & \overline{x}_1 = 21.2, & s_1^2 = 1.68, \\ \textbf{Estación 2} & n_2 = 10, & \overline{x}_2 = 19.2, & s_2^2 = 0.81. \end{array}$$

a) Verifique si las varianzas de las temperaturas pronosticadas son distintas. (Tome $\alpha = 0.1$);

b) Empleando el resultado anterior, pruebe si las dos estaciones pronostican iguales temperaturas, a un nivel de significación de 0.1.

27. Se van a probar dos medicamentos A y B contra una enfermedad. Para ésto se trataron 21 ratones enfermos con A y otros 21 con B. El número medio de horas que sobreviven con A es 1200 y el número medio con B es 1225. Suponiendo normalidad en ambos casos:

a) Se puede aceptar igualdad de varianzas si se sabe que $\sum(x_i - \overline{x})^2 = 9800$ y $\sum(y_i - \overline{y})^2 = 3500$? (Tome $\alpha = 0.05$);

b) ¿Es más efectivo el medicamento B? Plantee el contraste adecuado para estudiar esto con un nivel de significación del 5 %.

28. Una compañía petrolera está considerando la posibilidad de introducir un aditivo en su gasolina, esperando incrementar el kilometraje medio por litro. Los ingenieros del grupo de investigación probaron 10 autos con la gasolina habitual y otros 10 autos con la gasolina con el aditivo. El resumen de los resultados es:

$$\begin{array}{rll} \textbf{Kilometraje sin aditivo:} & \overline{x} = 14.2, & s_x = 3.24, \\ \textbf{Kilometraje con aditivo:} & \overline{y} = 15.4, & s_y = 5.56. \end{array}$$

a) ¿Es razonable suponer que las varianzas son iguales? Use $\alpha = 0.1$;

b) ¿La introducción del aditivo incrementa el kilometraje medio por litro?

Pruebas sobre la diferencia de dos proporciones

29. En una investigación en Guayaquil, 264 de 1140 mujeres y 214 de 1200 hombres respondieron que ellos no ingieren alcohol. Realice la prueba de hipótesis apropiada para asegurar si los datos muestrales proveen evidencia para afirmar que las mujeres guayaquileñas se abstienen de tomar alcohol en una proporción mayor que los hombres guayaquileños. En la resolución:

a) Especifique las hipótesis nula y alternativa, a la vez en símbolos y en palabras;

b) Calcule el estadístico de prueba;

c) Compare con el valor de la tabla y saque su conclusión.

30. Un economista del Instituto Nacional de Censos desea conocer si las tasas de desocupación urbanas son iguales en las dos principales ciudades del país. Con base en sendas muestras de 500 personas en cada una de las ciudades, el economista encontró 46 personas desocupadas en una ciudad y 35 en la otra. A un nivel de significación del 3 %, ¿puede suponerse que las tasas de desempleo en las dos ciudades son diferentes?

31. De una muestra de 300 televidentes escogidos al azar, que tenían sus televisores encendidos en la hora del noticiero, 50 indicaron que tenían sintonizado el canal 4 y 70 que sintonizaron el canal 8.

a) ¿Puede afirmarse que los dos canales tienen igual nivel de sintonía a la hora del noticiero? (Utilice un $\alpha = 2\,\%$);

b) Halle el nivel de significación de la prueba.

32. Dos empresas encuestadoras realizan investigaciones para determinar el porcentaje de personas que votarán a favor de una pregunta en un plebiscito. La primera empresa entrevistó a 1000 personas de las cuales 367 contestaron afirmativamente. La segunda empresa entrevistó a 300 personas y obtuvo 121 respuestas afirmativas.

   a) Puede considerarse que las dos empresas entregan resultados similares respecto del número de electores que votarán SI a la pregunta. Utilice $\alpha = 0.05$;
   b) Encuentre el nivel de significación de la prueba.

33. Se seleccionó, al azar, 500 usuarios de correo electrónico que trabajan en empresas públicas y resultó que 32 de ellos habían recibido virus informáticos a través del correo en el último año. Se realizó otro muestreo independiente eligiendo, al azar, 300 usuarios que trabajan en empresas privadas, resultando que 9 de ellos habían tenido este mismo problema.

   a) ¿Proporcionan estos datos suficiente evidencia estadística, al nivel 0.04, a favor de la hipótesis de que la incidencia de los virus es mayor en las empresas públicas?;
   b) El $p-$valor del contraste, ¿es mayor o menor que 0.01?

34. Un estudio indicaba que las hijas de madres fumadoras durante el embarazo tienen mayor probabilidad de ser ellas mismas fumadoras. El estudio se llevó a cabo con niñas, preguntándo si ellas habían fumado el último año y a la vez se consultó a las madres si ellas había fumado cuando estaban embarazadas. Solo el 4 % de las hijas, de 200 madres que no fumaron cuando estaban embarazadas, habían fumado el último año, comparado con el 26 % de las hijas, de 500 madres que habían fumado, también lo habían hecho. En la investigación se propone que el hecho que una madre haya fumado en estado de embarazo, aumenta la proporción de hijas fumadoras en un 20 %.

   a) ¿Qué hipótesis necesita usted para determinar si la diferencia indicada es estadísticamente significativa?;
   b) Realice una prueba de hipótesis para establecer si la propuesta de la investigación sobre el incremento en la proporción es plausible;
   c) ¿Establece el estudio que si una madre embarazada fuma, hay una mayor tendencia a fumar por parte de las hijas? Explique.

35. Se desea comparar la proporción de viviendas con servicio de alcantarillado en las áreas urbana y rural de Pastaza. Se hizo un muestreo en las dos áreas con los siguientes resultados:

   - Zona rural: De 500 viviendas elegidas al azar, 230 disponían de alcantarillado.
   - Zona urbana: De 1000 viviendas elegidas al azar, 680 disponían de alcantarillado.

   ¿Hay suficiente evidencia para concluir, con un nivel de significación del 3 %, que en Pastaza, la proporción de viviendas con alcantarillado en la zona urbana excede en más del 15 % a la proporción de viviendas con alcantarillado en la zona rural?

36. Para estudiar el efecto de una nueva terapia sobre el cáncer de seno se tomaron dos muestras, una de 300 pacientes que no recibieron la terapia y otra de 200 que si lo hicieron. De las que no recibieron, 171 pacientes murieron y de las que si recibieron, 66 murieron. Un tratamiento se considera efectivo si rebaja en más del 18 % el porcentaje de los pacientes fallecidos. A un nivel de significación de 0.05, ¿es efectivo el nuevo tratamiento?

Capítulo 9

Pruebas de Hipótesis No Paramétricas

KISS – Keep It Simple, but Scientific
Emanuel Parzen

En las pruebas de hipótesis que utilizan las distribuciones normal, t o F, se supone que la ley de probabilidad de la población de la cual se extrae la muestra tiene determinada forma y que sus parámetros verifican ciertas condiciones de manera que el estadístico muestral correspondiente tiene una distribución de probabilidad conocida, por lo que se llaman pruebas paramétricas. Por ejemplo, para la aplicación de la prueba t, se debe suponer que la población es normal. Además, para la aplicación de las pruebas paramétricas se requiere que el nivel de las mediciones sea al menos de escala de intervalos.

Sin embargo, existen muchas aplicaciones en las ciencias y la ingeniería donde no es posible conocer las distribuciones de las poblaciones de las que se extraen las muestras o los datos se reportan como valores en escala ordinal. En estos casos, se utilizan métodos alternativos equivalentes a los paramétricos, denominados *métodos no paramétricos* o *de distribución libre.*

Con frecuencia se utilizan las pruebas no paramétricas cuando se tratan de inferencias con muestras pequeñas y distribución desconocida de la población, ya que en estos casos no se puede utilizar el Teorema del Límite Central. La aplicación de los métodos no paramétricos no requiere conocimientos matemáticos avanzados, debido a que la tarea matemática consiste en ordenar por rangos los datos observados.

Si se verifican las condiciones exigidas para el uso de una prueba paramétrica, entonces, es siempre preferible utilizar ésta y no su equivalente no paramétrico. Ello se debe a que si se utiliza el mismo nivel de significación en ambas pruebas, la potencia de una prueba no paramétrica es siempre menor a la de su equivalente paramétrico. Por otro lado, con los métodos no paramétricos se pierde gran cantidad de información al no operar explícitamente con los valores sino con sus rangos.

Ventajas y desventajas de los métodos no paramétricos

Las pruebas no paramétricas tienen varias ventajas sobre las pruebas paramétricas:

1. Por lo general, son fáciles de usar y entender.
2. Eliminan la necesidad de suposiciones restrictivas de las pruebas paramétricas.

3. Se pueden usar con muestras pequeñas.
4. Se pueden usar con datos cualitativos.

También, las pruebas no paramétricas tienen desventajas:

1. A veces, ignoran, desperdician o pierden información.
2. No son tan eficientes como las paramétricas.
3. Llevan a una mayor probabilidad de no rechazar una hipótesis nula falsa (error de tipo II).

Las pruebas no paramétricas se dividen en dos grupos: sobre una sola muestra y sobre varias muestras. También, hay un grupo de pruebas que se basan en la ley de distribución ji cuadrado.

En este capítulo examinaremos un conjunto básico de pruebas no paramétricas, que permiten contrastar hipótesis sobre la independencia de variables, el ajuste a una ley de distribución dada, la aleatoriedad de las observaciones, entre otras. En la primera parte examinaremos aquellas pruebas que emplean la ley ji-cuadrado, mientras que la segunda parte dedicaremos a otro tipo de pruebas.

9.1. Pruebas χ^2 de bondad de ajuste a una ley

En esta sección analizaremos procedimientos cuyo objetivo es determinar si un conjunto de observaciones sigue cierto esquema probabilístico. Estos constrastes comparan las frecuencias observadas con las frecuencias teóricas del modelo probabilístico con que se coteja a través de un estadístico de prueba que sigue una ley χ^2.

9.1.1. Prueba sobre los parámetros en un experimento multinomial

Supongamos que en una investigación se establecen ciertas categorías en las cuales se divide la población; por ejemplo, el nivel socioeconómico, el nivel de instrucción, etc. Cada observación puede pertenecer solo a una de las k categorías prefijadas.

Si en la investigación se realizan n observaciones: n_1 de ellas están en la primera, n_2 están en la segunda, ..., n_k están en la $k-$ésima categoría; y se cumple que $n_1 + n_2 + \cdots + n_k = n$. Una investigación que presenta esta característica se denomina *experimento multinomial*, cuyas particularidades son:

1. El experimento consta de n ensayos idénticos e independientes unos de otros.
2. El resultado de cada ensayo se localiza en una y solo una de las categorías.
3. La probabilidad de que un ensayo se localice en la $i-$ésima categoría es p_i, $(i = 1, 2, \ldots, k)$; y se cumple que $p_1 + p_2 + \cdots + p_k = 1$.

Esto se resume en una tabla como la siguiente:

Categoría	C_1	C_2	$\cdots$	C_k	Total
Frecuencia	n_1	n_2	$\cdots$	n_k	n
Probabilidad	p_1	p_2	$\cdots$	p_k	1

Interesa conocer si el número de observaciones que se ubican en cada categoría se ajusta a un esquema de probabilidad dado; es decir, si las probabilidades de pertenencia a cada grupo tienen valores específicos: $p_1 = p_{10}$, $p_2 = p_{20}$, ..., $p_k = p_{k0}$. Entonces la prueba queda así:

1. *Hipótesis Nula.* H_0: $p_1 = p_{10}$, $p_2 = p_{20}$, ..., $p_k = p_{k0}$.
2. *Hipótesis Alternativa.* H_1: Al menos uno de los p_i es distinto de p_{i0}, $i = 1, 2, \ldots, k$.
3. *Estadístico de Prueba.* $\chi^2_{obs} = \sum_{i=1}^{k} \frac{(n_i - n\, p_{i0})^2}{n\, p_{i0}}$.
4. *Región de Rechazo.* Se rechaza H_0 cuando $\chi^2_{obs} > \chi^2_\alpha(k-1)$.

Ejemplo. En un periodo de un año, se registraron 100 nacimientos de gemelos en la Maternidad Isidro Ayora de Quito. La distribución de los recién nacidos, de acuerdo al sexo, es la siguiente:

2 varones	2 mujeres	1 varón, 1 mujer
29	38	33

Se supone que los datos están distribuidos según una ley trinomial de parámetros $(100; p_1, p_2, p_3)$ con $p_1 = p_2 = \frac{1}{4}$ y $p_3 = \frac{1}{2}$. Probar la hipótesis a un nivel de significación del 5 %.

Solución: Se tiene $n_1 = 29$, $n_2 = 38$, $n_3 = 33$ y $n = 100$, con esos valores la prueba es

1. *Hipótesis Nula.* H_0: $p_1 = \frac{1}{4}$, $p_2 = \frac{1}{4}$, $p_3 = \frac{1}{2}$.
2. *Hipótesis Alternativa.* H_1: Al menos uno de los p_i es distinto del enunciado.
3. *Estadístico de Prueba.*

$$\begin{aligned} \chi^2_{obs} &= \sum_{i=1}^{k} \frac{(n_i - n\, p_{i0})^2}{n\, p_{i0}} = \frac{\left(29 - \frac{1}{4} \times 100\right)^2}{\frac{1}{4} \times 100} + \frac{\left(38 - \frac{1}{4} \times 100\right)^2}{\frac{1}{4} \times 100} + \frac{\left(33 - \frac{1}{2} \times 100\right)^2}{\frac{1}{2} \times 100} \\ &= 13.18. \end{aligned}$$

4. *Región de Rechazo.* Tenemos que $\chi^2_{0.05}(2) = 5.99$ y la región crítica es $\chi^2_{obs} > 5.99$.
5. *Decisión.* Como $13.18 > 5.99$, se rechaza H_0, o sea, los nacimientos de los gemelos no siguen la ley indicada. ◀

Se recomienda que el lector cambie las probabilidades de pertenencia a cada grupo, de manera que se acepte la hipótesis nula.

9.1.2. Prueba de bondad de ajuste a una ley dada

Este caso es una generalización del anterior. Ahora, supondremos que no podemos asignar de antemano las probabilidades de pertenencia a cada grupo, sino que se las debe calcular a partir de una ley de distribución teórica (uniforme, normal, Poisson, etc.), cuyos parámetros son conocidos o deben ser estimados a partir de las observaciones de la muestra.

Disponemos de un conjunto de n observaciones, que se supone siguen una ley de probabilidad dada y que están agrupadas en k clases o categorías.

Por otro lado, sea X una variable aleatoria que sigue la indicada ley probabilidad, que toma valores x_1, x_2, ...; entonces, $\Pr(Y = x_i) = p_i$.

Por ejemplo, X podría seguir una ley de Poisson,

$$\Pr(X = x_i) = p_i = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!},$$

donde el parámetro λ puede estar previamente especificado o debe ser estimado.

A partir de las probabilidades teóricas se calculan las frecuencias esperadas de cada clase como $e_i = np_i$.

Cuando alguna clase tiene una frecuencia observada menor que 5 se la puede unir con alguna clase adyacente y sumar las probabilidades correspondientes. Luego de agrupar las observaciones que lo ameriten, se dispondrá de una tabla de frecuencias con k clases, como la siguiente:

Grupo o clase	**Observación**	**Frecuencia observada (n_i)**	**Frecuencia esperada (e_i)**
1	x_1	n_1	$e_1 = np_1$
2	x_2	n_2	$e_2 = np_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	x_k	n_k	$e_k = np_k$
	Total	n	n

El estadístico de prueba, para comprobar si los datos siguen una ley específica, es

$$\chi^2_{obs} = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i},$$

que sigue aproximadamente una distribución χ^2 con $[(k-1)-(\textit{número de parámetros estimados})]$ grados de libertad. Así, si se supone una ley de Poisson, de parámetro λ conocido, entonces χ^2_{obs} sigue una ley $\chi^2(k-1)$; pero si se estima el parámetro λ, el estadístico χ^2_{obs} sigue una distribución $\chi^2(k-2)$.

La prueba de hipótesis queda como sigue:

1. *Hipótesis Nula.* H_0: Los datos siguen una ley $\mathcal{L}(p)$ dada.
2. *Hipótesis Alternativa.* H_1: Los datos no siguen la ley $\mathcal{L}(p)$ dada.
3. *Estadístico de Prueba.* $\chi^2_{obs} = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}$,

   donde k es el número de clases que se forman luego de agrupar los datos.
4. *Región de Rechazo.* Se rechaza H_0 si $\chi^2_{obs} > \chi^2_{\alpha}(k-1-l)$, donde l = número de parámetros estimados a partir de la muestra.

Ejemplos

1. En una agencia bancaria hay cinco cajas para atender a los clientes. Un día, el encargado de la agencia contabilizó el número de clientes que escogía cada una de las cajas, obteniendo:

Caja	1	2	3	4	5	Total
Frecuencia	34	54	39	48	45	220

   De acuerdo a estos resultados, ¿se puede concluir que hay preferencia por alguna de las cajas?

Solución: La hipótesis nula es suponer que no hay preferencia por una caja particular o, lo que es lo mismo, que los datos siguen una ley uniforme con

$$p_i = \Pr(Y = i) = \frac{1}{5}, \quad i = 1, \ldots, 5.$$

Entonces, las frecuencias esperadas son:

$$e_i = n\,p_i = 220 \times \frac{1}{5} = 44, \quad i = 1, \ldots, 5.$$

Notemos que aquí no es necesario estimar parámetro alguno; de manera que la prueba es:

1. *Hipótesis Nula.* H_0: Los datos siguen una ley uniforme.
2. *Hipótesis Alternativa.* H_1: Los datos no siguen una ley uniforme.
3. *Estadístico de Prueba.*

$$\begin{aligned}\chi^2_{obs} &= \sum_{i=1}^{5} \frac{(n_i - e_i)^2}{e_i} \\ &= \frac{(34-44)^2}{44} + \frac{(54-44)^2}{44} + \frac{(39-44)^2}{44} + \frac{(48-44)^2}{44} + \frac{(45-44)^2}{44} = 5.5.\end{aligned}$$

4. *Región de Rechazo.* Como $\chi^2_{0.05}(5-1) = 9.49$, la región es $\chi^2_{obs} > 9.49$.
5. *Decisión.* Aquí, χ^2_{obs} no está en la región de rechazo; de manera que se puede decir que no hay preferencia por ninguna caja.

2. En una ensambladora de carros se registró el número de defectos por unidad en una muestra de 100 unidades que se inspeccionaron durante una semana dada, dando la siguiente distribución de frecuencias:

Número de defectos	0	1	2	3	4
Número de carros	63	20	8	5	4

Contrastar la hipótesis de que los datos se ajustan a una distribución de Poisson.

Solución: El estimador del parámetro λ de la ley de Poisson es el promedio:

$$\widehat{\lambda} = \overline{x} = \frac{0 \times 63 + 1 \times 20 + 2 \times 8 + 3 \times 5 + 4 \times 4}{100} = 0.67.$$

Calculemos las probabilidades de pertenencia a cada grupo, según la ley $\mathcal{P}(0.67)$:

$$\begin{aligned}p_0 &= \Pr(X = 0) = \frac{e^{-0.67}(0.67)^0}{0!} = 0.512, \\ p_1 &= \Pr(X = 1) = \frac{e^{-0.67}(0.67)^1}{1!} = 0.343, \\ p_2 &= \Pr(X = 2) = \frac{e^{-0.67}(0.67)^2}{2!} = 0.115, \\ p_3 &= \Pr(X = 3) = \frac{e^{-0.67}(0.67)^3}{3!} = 0.026, \\ p_4 &= \Pr(X = 4) = \frac{e^{-0.67}(0.67)^4}{4!} = 0.004.\end{aligned}$$

Las frecuencias esperadas son

$$e_0 = 100 \times 0.512 = 51.2, \qquad e_1 = 100 \times 0.343 = 34.3,$$
$$e_2 = 100 \times 0.115 = 11.5, \qquad e_3 = 100 \times 0.026 = 2.6,$$
$$e_4 = 100 \times 0.004 = 0.4.$$

Calculemos el estadístico de prueba:

$$\begin{aligned}\chi^2_{obs} &= \sum_{i=0}^{4} \frac{(n_i - e_i)^2}{e_i} \\ &= \frac{(63-51.2)^2}{51.2} + \frac{(20-34.3)^2}{34.3} + \frac{(8-11.5)^2}{11.5} + \frac{(5-2.6)^2}{2.6} + \frac{(4-0.4)^2}{0.4} = 41.72.\end{aligned}$$

La prueba estadística es:

1. *Hipótesis Nula.* H_0: Los datos siguen una ley de Poisson $\mathcal{P}(0.67)$.
2. *Hipótesis Alternativa.* H_1: Los datos no siguen una ley de Poisson $\mathcal{P}(0.67)$.
3. *Estadístico de Prueba.* $\chi^2_{obs} = 41.72$.
4. *Región de Rechazo.* Como $\chi^2_{0.05}(5-1-1) = 7.81$ y la región es $\chi^2_{obs} > 7.81$.
5. *Decisión.* Se cumple que $41.72 > 7.81$; entonces, se rechaza la hipótesis nula. El número de defectos no siguen la ley de Poisson $\mathcal{P}(0.67)$.

3. Se midió la duración de trabajo de 200 elementos electrónicos, según se resume en la tabla. La primera columna muestra los intervalos de tiempo (en horas) y en la segunda está la cantidad de elementos que han trabajado el tiempo entre los límites del correspondiente intervalo.

Tiempo de duración (hr)	**Frecuencia** (n_i)	**Tiempo de duración (hr)**	**Frecuencia** (n_i)
0 - 5	133	15 - 20	4
5 - 10	45	20 - 25	2
10 - 15	15	25 - 30	1

Para el nivel de significación 0.01, verificar la hipótesis de que el tiempo medio de trabajo de los elementos está distribuido según una ley exponencial.

Solución: El parámetro λ, que sigue la ley exponencial, se estima por $\widehat{\lambda} = \frac{1}{\overline{x}} = 0.2$.

Calculemos las probabilidades de que la variable aleatoria tome valores en cada intervalo:

$$p_1 = \Pr(0 \leq X < 5) = e^{-0.2\times 0} - e^{-0.2\times 5} = 0.6321.$$

Análogamente, se obtienen

$$p_2 = 0.2326, \quad p_3 = 0.0855, \quad p_4 = 0.0315, \quad p_5 = 0.0116, \quad p_6 = 0.0043.$$

Con ésto, se encuentran las frecuencias esperadas, $e_i = np_i$:

$$e_1 = 126.42, \quad e_2 = 46.52, \quad e_3 = 17.10, \quad e_4 = 6.30, \quad e_5 = 2.32, \quad e_6 = 0.86.$$

Como las tres últimas frecuencias son pequeñas, se las puede agrupar en una sola clase, obteniéndose $k = 4$ clases. De esta manera, se dispone de la siguiente tabla que muestra el grupo, la frecuencia observada y la frecuencia teórica.

i	n_i	e_i
1	133	126.42
2	45	46.52
3	15	17.10
4	7	9.48
Total	200	

Por lo tanto, la prueba es:

1. *Hipótesis Nula.* H_0: Las observaciones siguen una ley exponencial $\mathcal{E}(0.2)$.
2. *Hipótesis Alternativa.* H_1: Los datos no siguen una ley exponencial $\mathcal{E}(0.2)$.
3. *Estadístico de Prueba.*

$$\begin{aligned}\chi^2_{obs} &= \sum_{i=1}^{4}\frac{(n_i-e_i)^2}{e_i} = \frac{(133-126.42)^2}{126.42}+\frac{(45-46.52)^2}{46.52}+\frac{(15-17.10)^2}{17.10}+\frac{(7-9.48)^2}{9.48}\\ &= 1.299.\end{aligned}$$

4. *Región de Rechazo.* Como previamente se estimó λ, $l = 1$ y se deberá examinar la ley χ^2 con $(4-1-1) = 2$ grados de libertad, cuyo valor es $\chi^2_{0.01}(2) = 9.21$, se define la región de rechazo $\chi^2_{obs} > 9.21$.
5. *Decisión.* Puesto que $1.299 < 9.21$, no hay razón para rechazar la hipótesis de una ley exponencial $\mathcal{E}(0.2)$ para los datos.

4. Se midió la estatura de un grupo de niños de 12 años de edad, con los resultados:

Estatura	**Número de niños**
135 - 140	6
140 - 145	18
145 - 150	32
150 - 155	61
155 - 160	22
160 - 165	5

Verificar si los datos de esta muestra provienen de una distribución normal.

Solución: En primer lugar se deben estimar los parámetros de la ley normal:

$$\widehat{\mu} = \overline{x} = 150.625, \quad \widehat{\sigma^2} = 31.425.$$

Las probabilidades de pertenencia a cada intervalo, según una ley $\mathcal{N}(150.625; 31.425)$ son:

$$p_1 = 0.026, \quad p_2 = 0.129, \quad p_3 = 0.298, \quad p_4 = 0.327, \quad p_5 = 0.170, \quad p_6 = 0.042.$$

Y las frecuencias esperadas son:

$$e_1 = 4.18, \quad e_2 = 18.55, \quad e_3 = 42.88, \quad e_4 = 47.06, \quad e_5 = 24.77, \quad e_6 = 6.48.$$

Entonces, la prueba es:

1. *Hipótesis Nula.* H_0: Las observaciones siguen una ley normal $\mathcal{N}(150.625; 31.425)$.
2. *Hipótesis Alternativa.* H_1: Los datos no siguen una ley normal $\mathcal{N}(150.625; 31.425)$.
3. *Estadístico de Prueba.* $\chi^2_{obs} = 8.349$.
4. *Región de Rechazo.* Como se estimaron dos parámetros, entonces $l = 2$, el estadístico de prueba se debe comparar con $\chi^2_{0.05}(6-1-2) = 7.81$. La región crítica es $\chi^2_{obs} > 7.81$.
5. *Decisión.* Como $8.349 > 7.81$, se rechaza H_0; por tanto, las estaturas de los niños de la muestra no siguen la ley normal enunciada. ◄

A continuación se presenta un resumen de los estimadores de los parámetros de las leyes más comunmente empleadas.

Ley	Estimador
Binomial $\mathbf{Bin}(n,p)$	$\widehat{p} = \frac{y}{n}$
Hipergeométrica $\mathcal{H}(N,p,r)$	$\widehat{p} = \frac{\overline{x}}{r}$
Poisson $\mathcal{P}(\lambda)$	$\widehat{\lambda} = \overline{x}$
Geométrica $\mathcal{G}(p)$	$\widehat{p} = \frac{1}{\overline{x}}$
Uniforme $\mathcal{U}(a,b)$	$\widehat{a} = \overline{x} - \sqrt{3}s$ $\widehat{b} = \overline{x} + \sqrt{3}s$
Exponencial $\mathcal{E}(\lambda)$	$\widehat{\lambda} = \frac{1}{\overline{x}}$
Normal $\mathcal{N}(\mu,\sigma^2)$	$\widehat{\mu} = \overline{x}$ $\widehat{\sigma}^2 = s^2$

9.2. Pruebas sobre tablas de contingencia

Cuando tenemos la información de 2 variables de tipo cualitativo, se la resume en una *tabla de contingencia*, que es una tabla de frecuencias de doble entrada, donde en las filas se ponen las modalidades de una variable, y en las columnas las modalidades de la otra; en las celdas resultantes del cruce de las filas y las columnas se coloca el número de elementos que presentan ambas modalidades.

Si se tiene información de N elementos acerca de las variables A y B, de tal forma que presentan r y c modalidades respectivamente, la tabla de contingencia $r \times c$ (r filas y c columnas) es de la forma:

	Variable B					
Variable A	B_1	$\cdots$	B_j	$\cdots$	B_c	Total
A_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1c}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{ic}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_r	n_{r1}	$\cdots$	n_{rj}	$\cdots$	n_{rc}	$n_{r\cdot}$
Total	$n_{\cdot 1}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot c}$	N

donde,

$$n_{i\cdot} = \sum_{j=1}^{c} n_{ij} \quad \text{(total de la frecuencia de la fila } i\text{)},$$

$$n_{\cdot j} = \sum_{i=1}^{r} n_{ij} \quad \text{(total de la frecuencia de la columna } j\text{)}.$$

Por ejemplo, los jefes de familia que viven en una ciudad pueden clasificarse en dueños o arrendatarios de la casa donde residen y según su nivel de instrucción en primario, secundario y superior; es decir, se los podría clasificar en una tabla de contingencia 2×3.

9.2.1. Prueba de independencia

Considérese las probabilidades $p_{i\cdot}$, de encontrarse en la fila i; $p_{\cdot j}$, de encontrarse en la columna j; y p_{ij} la probabilidad de encontrarse en la celda (i,j).

Debe cumplirse que $\sum_{i=1}^{r} p_{i\cdot} = 1$ y $\sum_{j=1}^{c} p_{\cdot j} = 1$, y pueden ser estimadas por

$$\widehat{p}_{i\cdot} = \frac{n_{i\cdot}}{N}, \quad i = 1, \ldots, r,$$
$$\widehat{p}_{\cdot j} = \frac{n_{\cdot j}}{N}, \quad j = 1, \ldots, c.$$

Bajo la hipótesis de independencia entre filas y columnas, se tiene que la frecuencia esperada en la celda ubicada en la i−ésima fila y j−ésima columna es

$$e_{ij} = N\widehat{p}_{i\cdot}\, \widehat{p}_{\cdot j} = \frac{n_{i\cdot}\, n_{\cdot j}}{N}.$$

El estadístico que permite probar la hipótesis de independencia es

$$\chi^2_{obs} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}},$$

que sigue aproximadamente una distribución χ^2 con $[(r-1)(c-1)]$ grados de libertad.

Entonces, la prueba para la independencia es la siguiente:

1. *Hipótesis Nula.* H_0: $p_{ij} = p_{i\cdot}\, p_{\cdot j}$ (la variable A es independiente de la variable B).
2. *Hipótesis Alternativa.* H_1: $p_{ij} \neq p_{i\cdot}\, p_{\cdot j}$, para al menos una celda de la tabla (la variables A y B no son independientes).
3. *Estadístico de Prueba.* $\chi^2_{obs} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$.
4. *Región de Rechazo.* La hipótesis de independencia se rechaza si $\chi^2_{obs} > \chi^2_{\alpha}[(r-1)(c-1)]$.

Ejemplo. En una investigación se desea revelar si existe relación entre el consumo de combustible y el origen de los carros que circulan por la ciudad.

	Origen			
Consumo	EE.UU.	Europa	Japón	Total
Bajo	76	56	70	202
Alto	160	14	9	183
Total	236	70	79	385

Al nivel 5 %, verificar si las dos variables están asociadas.

Solución: Calculemos las frecuencias esperadas, e_{ij}:

$$e_{11} = \frac{n_{1\cdot}\, n_{\cdot 1}}{N} = \frac{202 \times 236}{385} = 123.82, \qquad e_{12} = \frac{n_{1\cdot}\, n_{\cdot 2}}{N} = \frac{202 \times 70}{385} = 36.72,$$
$$e_{13} = \frac{n_{1\cdot}\, n_{\cdot 3}}{N} = \frac{202 \times 79}{385} = 41.45, \qquad e_{21} = \frac{n_{2\cdot}\, n_{\cdot 1}}{N} = \frac{183 \times 236}{385} = 112.18,$$
$$e_{22} = \frac{n_{2\cdot}\, n_{\cdot 2}}{N} = \frac{183 \times 70}{385} = 33.27, \qquad e_{23} = \frac{n_{2\cdot}\, n_{\cdot 3}}{N} = \frac{183 \times 79}{385} = 37.55.$$

Con ellas formemos la tabla de frecuencias esperadas:

Consumo	**Origen**		
	EE.UU.	Europa	Japón
Bajo	123.82	36.72	41.45
Alto	112.18	33.27	37.55

Entonces, la prueba queda como

1. *Hipótesis Nula.* H_0: El origen del carro y el consumo de combustible son independientes.
2. *Hipótesis Alternativa.* H_1: El origen del carro y el consumo de combustible están relacionados.
3. *Estadístico de Prueba.*

$$\begin{aligned}\chi^2_{obs} &= \frac{(76-123.82)^2}{123.82}+\frac{(56-36.72)^2}{36.72}+\frac{(70-41.45)^2}{41.45}\\ &\quad+\frac{(160-112.18)^2}{112.18}+\frac{(14-33.27)^2}{33.27}+\frac{(9-37.55)^2}{37.55}\\ &= 101.51.\end{aligned}$$

4. *Región de Rechazo.* El número de grados de libertad es $\nu = (2-1)(3-1) = 2$ y si $\alpha = 0.05$, se tiene que $\chi^2_{0.05}(2) = 5.99$. La región de rechazo es $\chi^2_{obs} > 5.99$.
5. *Decisión.* Como $101.51 > 5.99$, se rechaza la hipótesis nula. Concluimos que hay relación entre el consumo de combustible y el origen de los carros.

Tablas de contingencia de 2×2

Para el caso especial de la tabla de contingencia de 2×2, se ha desarrollado una fórmula de cálculo de χ^2_{obs} que no requiere la determinación de las frecuencias esperadas. Supongamos que se dispone de una tabla de contingencia de dos variables, cada una de las cuales solo pueden tomar dos valores, como la siguiente:

Variable A	Variable B		Total
	B_1	B_2	
A_1	a	b	$a+b$
A_2	c	d	$c+d$
Total	$a+c$	$b+d$	n

El estadístico de prueba se calcula como

$$\chi^2_{obs} = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)},$$

que sigue una ley χ^2 a 1 grado de libertad.

El resto de la prueba de hipótesis queda igual a lo enunciado anteriormente.

Ejemplo. El siguiente cuadro muestra la reacción de los espectadores ante un comercial de televisión, clasificados por sexo.

Reacción	Sexo		Total
	Hombres	Mujeres	
Desfavorable	10	5	15
Favorable	3	7	10
Total	13	12	25

¿Depende la aceptación del comercial del sexo del televidente?

Solución: La prueba de hipótesis queda de la siguiente manera:

1. *Hipótesis Nula.* H_0: La reacción ante el comercial es independiente del sexo.
2. *Hipótesis Alternativa.* H_1: La reacción ante el comercial y el sexo están relacionados.
3. *Estadístico de Prueba.* $\chi^2_{obs} = \dfrac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} = \dfrac{25(10\times 7 - 5\times 3)^2}{15\times 10\times 13\times 12} = 3.23.$
4. *Región de Rechazo.* Si escogemos $\alpha = 0.05$, vemos que $\chi^2_{0.05}(1) = 3.84$. La región de rechazo es $\chi^2_{obs} > 3.84$.
5. *Decisión.* Como $3.23 < 3.81$, no se rechaza la hipótesis nula. Se concluye que hay independencia entre el sexo del espectador y la aceptación del comercial.

Se sugiere que el lector realice el ejercicio mediante el cálculo de las frecuencias esperadas.

9.2.2. Prueba de homogeneidad

Se dice que una muestra es *homogénea* si todas sus observaciones han sido generadas por el mismo modelo de distribución de probabilidad o pertenecen a una misma población. La metodología de las pruebas de independencia también puede aplicarse para averiguar si existe homogeneidad entre dos o más muestras independientes.

Así, la prueba estadística es la siguiente:

1. *Hipótesis Nula.* H_0: Las muestras provienen de una misma población (las muestras son homogéneas).
2. *Hipótesis Alternativa.* H_1: Las muestras no provienen de una misma población (las muestras son heterogéneas).
3. *Estadístico de Prueba.* $\chi^2_{obs} = \sum_{i=1}^{r}\sum_{j=1}^{c}\dfrac{(n_{ij}-e_{ij})^2}{e_{ij}}$.
4. *Región de Rechazo.* La hipótesis de independencia se rechaza si $\chi^2_{obs} > \chi^2_{\alpha}[(r-1)(c-1)]$.

Ejemplo. En una Facultad se clasificó a las notas obtenidas por sus alumnos, luego de rendir el mismo examen de Física, como bajas, medias y altas. También, se registró el profesor que dictaba la materia, obteniendo:

Profesor	Calificación			Total
	Baja	Media	Alta	
A	12	23	7	42
B	25	17	4	46
Total	37	40	11	88

¿Se puede decir que la diferencia en las notas es debida al profesor?

Solución: La tabla de frecuencias esperadas es:

	Calificación		
Profesor	Baja	Media	Alta
A	17.7	19.1	5.3
B	19.3	20.9	5.8

La prueba estadística es:

1. *Hipótesis Nula.* H_0: Las diferencias en las notas no son debidas al profesor de la materia.
2. *Hipótesis Alternativa.* H_1: Las diferencias en las notas se deben al profesor.
3. *Estadístico de Prueba.* $\chi^2_{obs} = \dfrac{(12-17.7)^2}{17.7} + \dfrac{(23-19.1)^2}{19.1} + \cdots + \dfrac{(4-5.8)^2}{5.8} = 6.12.$
4. *Región de Rechazo.* Se tiene que $\chi^2_{0.05}(2) = 5.99$ y la región es $\chi^2_{obs} > 5.99$.
5. *Decisión.* Como $6.12 > 5.99$, se rechaza H_0; es decir, el profesor de la materia si influye en las notas obtenidas en su materia.

9.3. Ejercicios

Pruebas de bondad de ajuste a una ley

1. En un cruce de carreteras los autos pueden girar a la izquierda, a la derecha o seguir de frente. Se supone que la mitad de los autos seguirán de frente, la una cuarta parte irá a la izquierda y la cuarta parte restante a la derecha. Se realizó un conteo de los autos según la dirección que ellos siguen:

	De frente	A la izquierda	A la derecha
Frecuencia	29	14	9

Pruebe la hipótesis indicada a un nivel de significación de 0.1.

2. Cuando el naturalista francés del siglo XVIII Georges Louis Buffon realizó 4040 lanzamientos de una moneda observó 2048 caras. ¿Concuerdan estos datos con la hipótesis de que la moneda es simétrica?

3. Una zona de Mindo es el hábitat natural de tres especies de colibríes. Se cree (hipótesis nula) que una quinta parte de los colibríes pertenece a la primera especie, dos quintas partes a la segunda especie y otras dos quintas partes a la tercera. En una muestra aleatoria de 34 colibríes de la zona, se observaron 12 de la primera especie, 15 de la segunda y 7 de la tercera. ¿Hay suficiente evidencia estadística (al nivel 0.05) para aceptar la hipótesis propuesta?

4. Según los datos de un estudio exhaustivo de mercado que se realizó en la ciudad, las ventas de impresoras para computadoras personales de uso doméstico se dividen entre cuatro marcas (A, B, C y D) cuyos porcentajes del total de las ventas son 18 %, 22 %, 35 % y 25 %, respectivamente. Un año después, se quiere analizar de nuevo la situación pero se cree que no se debe repetir un estudio de mercado a gran escala. Se decide observar la marca adquirida por 200 compradores de impresoras elegidos al azar. obteniendo que de ellos 28 habían elegido la marca A, 48 la B, 77 la C y 47 la D. ¿Hay suficiente evidencia. al nivel 0.05, para afirmar que el reparto del mercado ya no es el mismo que el año anterior?

5. Dos empresas S y T especializadas en realizar sondeos políticos condujeron dos investigaciones sobre la intención de voto a un partido político. También, se disponen de los resultados de la elección llevada a cabo días después

Partido	Empresa S	Empresa T	Elección
A	416	188	32.9
B	62	22	5.1
C	193	83	15.9
F	139	76	12.4
M	95	36	10.6
V	151	71	14.0
Z	76	32	9.1
Total	1132	508	100.0

a) Compare los resultados de la elección con cada una de las investigaciones de las empresas;

b) ¿Se puede determinar cuál de las empresas dio el resultado más certero? Explique.

6. Cuando Gregor Mendel realizó sus experimentos de cruzamiento de guisantes, observó las frecuencias de varias semillas producidas por híbridos de guisantes amarillos lisos y guisantes verdes rugosos. Estos datos y sus respectivas probabilidades, según las predicciones de la teoría de la herencia de Mendel, se dan en la siguiente tabla.

Semillas	Frecuencia	Probabilidad
Amarilla y lisa	315	9/16
Amarilla y rugosa	101	3/16
Verde y lisa	108	3/16
Verde y rugosa	31	1/16
Total	$n = 555$	1

Para un nivel de significación del 1 %, examine si las frecuencias de los datos son consistentes con las probabilidades teóricas.

7. En el desarrollo decimal de π entre los 10 038 primeros dígitos aparecen 968 ceros, 1062 unos, 1021 dos, 974 tres, 1014 cuatros, 1046 cincos, 1021 seis, 970 sietes, 948 ochos y 1014 nueves.

a) ¿Se puede considerar que los dígitos aparecen uniformemente distribuidos, a un nivel de significación $\alpha = 0.05$?;

b) A qué nivel de significación rechazará la hipótesis?

8. Entre 2000 familias que tienen 2 hijos, 522 familias tienen dos varones, 471 dos mujeres y 1007 de los dos sexos. A un nivel de significación de 0.05, ¿se puede considerar que el número de hijos varones en las familias es una variable aleatoria binomial?

9. En un estudio a 107 familias que tienen 5 hijos, se contabilizó el número de hijas que tienen tales familias, con el siguiente resultado:

Número de hijas	0	1	2	3	4	5
Frecuencia	5	17	28	32	19	6

¿Puede afirmarse que, para las familias estudiadas, el número de hijas sigue una ley binomial? (Use $\alpha = 0.05$.)

10. Para estudiar la delincuencia en una ciudad se anotaron las denuncias por robo de automóviles recibidas en los últimos 575 días, obteniéndose los siguientes resultados:

No. denuncias	0	1	2	3	4	5
Frecuencia	230	210	90	35	8	2

Contraste la hipótesis de que los datos proceden de una distribución de Poisson con $\lambda = 1$. ($\alpha = 0.05$)

11. Se registró la cantidad de goles conseguidos en 149 partidos disputados por un equipo profesional de fútbol

No. de goles en un partido	0	1	2	3	4
Frecuencia	50	64	23	8	4

Verifique la hipótesis de que el número de goles por partido está distribuido según una ley de Poisson, para el nivel de significación 0.05.

12. En el transcurso de dos horas, el número de llamadas por minuto, solicitadas a una central telefónica fue:

No. llamadas/min	0	1	2	3	4	5	6
Frecuencia	6	18	32	35	17	10	2

¿Se puede aceptar que el número de llamadas por minuto sigue una distribución de Poisson?

13. En cierta región se registró la temperatura del aire durante 300 días. Las mediciones se resumen en la siguiente tabla (en la primera columna se indica el intervalo de temperatura en grados y en la segunda el número de días cuya temperatura media corresponde a ese intervalo). Verifique que la temperatura media está distribuida uniformemente a un nivel de significación de 0.05.

Rango de temperatura	Frecuencia	Rango de temperatura	Frecuencia
(-10) a (-5)	25	10 a 15	40
(-5) a 0	40	15 a 20	46
0 a 5	30	20 a 25	48
5 a 10	45	25 a 30	26

14. Un estudio realizado indica que el tiempo de espera en la cola de la caja de un banco se puede modelizar con una distribución exponencial de media 3 minutos. Para comprobar si este modelo sigue siendo válido, se tomó la siguiente muestra:

$$3 \quad 2 \quad 6 \quad 4 \quad 1 \quad 3 \quad 2 \quad 1 \quad 1 \quad 4.$$

Verifique si los datos proceden del modelo especificado en el estudio ($\alpha = 0.05$).

15. Se probaron 450 focos, registrándose el tiempo que duraron encendidos hasta que fallaron:

Tiempo	Número de focos	Tiempo	Número de focos
0 a 400	131	1600 a 2000	35
400 a 800	95	2000 a 2400	36
800 a 1200	76	2400 a 2800	21
1200 a 1600	56		

Para un nivel de significación de 0.01, verifíque la hipótesis de que el tiempo de duración de los focos está distribuida según una ley exponencial.

16. En la Facultad de Ciencias se quiere averiguar los conocimientos sobre Física que tienen los alumnos matriculados por primera vez en dicha Facultad, para lo cual se realizó el primer día de clase una prueba general. Los resultados correspondientes a una muestra de 211 alumnos se recogen en la siguiente tabla:

Puntuaciones	Frecuencia
50.5 – 55.5	4
55.5 – 60.5	17
60.5 – 65.5	45
65.5 – 70.5	67
70.5 – 75.5	53
75.5 – 80.5	15
80.5 – 85.5	10

Determine la normalidad de la variable que representa la puntuación obtenida. (Use $\alpha = 0.05$)

Pruebas sobre tablas de contingencia

17. En una investigación sobre el hábito de fumar por un grupo de estudiantes universitarios se obtuvo la siguiente tabla, en la cual se relaciona el sexo del entrevistado y si él o ella es un fumador.

	¿Fuma?	
Sexo	**SI**	**NO**
Masculino	26	10
Femenino	11	15

Pruebe si existe asociación entre el hábito de fumar y el sexo del investigado.

18. En un estudio médico a 300 pacientes que fueron operadas por cáncer de ovario, se clasificaron en quienes han sobrevivido 10 años después de la operación y quienes no lo hicieron, y el estado del tumor al momento de la operación.

Estado del cáncer	**¿Sobrevivió?**	
	No	Si
Temprano	32	127
Avanzado	118	23

¿La sobrevivencia de la paciente es independiente del estado del tumor el momento de la operación?

19. A fin de probar el supuesto de que una persona desempleada representa un alto riesgo crediticio, en un banco se realizó un estudio de 100 cuentas escogidas aleatoriamente con los siguientes resultados:

Situación actual del préstamo	**Situación laboral del cliente**	
	Empleado	Desempleado
En mora	16	10
No en mora	55	19

a) Calcule el valor de χ^2_{obs} para estos datos, usando la fórmula usual;

b) Calcule el valor de χ^2_{obs} para esta tabla mediante la fórmula alternativa;

c) ¿Cuál es el resultado de esta prueba al nivel de significación de 5 por ciento?

20. Se realizó un análisis de sangre de un grupo de 1000 habitantes (elegidos al azar) de una ciudad con la siguiente distribución, según el grupo sanguíneo y el factor Rh:

	O	A	B	AB
Rh+	357	317	81	39
Rh−	96	82	29	10

Según estos datos, ¿puede aceptarse la hipótesis de independencia del factor Rh del grupo sanguíneo? (Use $\alpha = 0.05$)

21. Se ejecutó un estudio sobre la utilización de ciertas fuentes de financiamiento externas para las pequeñas y medianas empresas (PYMES). Para ello, se seleccionó aleatoriamente 500 PYMES a nivel nacional. Las empresas se clasificaron según su tamaño en tres categorías (micros, pequeñas y medianas) y según hayan utilizado o no alguna fuente de financiación. Los datos obtenidos fueron:

	Con financiamiento	**Sin financiamiento**
Micros	115	325
Pequeñas	20	20
Medianas	15	5

a) ¿Existe alguna relación entre el tamaño de la empresa y el hecho de recurrir o no a fuentes de financiación? Utilice un nivel de significación del 10 %;

b) ¿Puede aceptarse, a un nivel de significación del 5 %, que un 20 % de las empresas «micros» utilizan fuentes de financiación frente a que la proporción sea mayor?

22. Se realizó un sondeo en la ciudad para determinar las posibles relaciones entre el nivel educativo (superior, medio o primario) de las personas y el consumo (bajo, medio o alto) de productos electrónicos. Los resultados, para 400 personas seleccionadas al azar, fueron:

Nivel	**Consumo**		
Educativo	Bajo	Medio	Alto
Superior	31	41	44
Medio	28	79	125
Primario	16	17	19

Contraste la independencia entre el nivel educativo y el consumo de productos electrónicos.

23. En una investigación sociológica a un grupo de personas casadas, se desea saber si el nivel socioeconómico de los encuestados incide sobre el éxito o fracaso de su matrimonio. Los resultados se dan en la siguiente tabla:

	NSE				
¿Fracasó?	I	II	III	IV	V
Si	28	62	79	181	124
No	127	230	443	850	582

¿Se puede concluir que la diferencia en el índice de fracaso se debe al nivel socioeconómico de los matrimonios?

24. El Consejo Directivo de una universidad quería determinar la opinión de diversos grupos en relación con el calendario docente propuesto. Una muestra aleatoria seleccionada entre 100 estudiantes, 50 empleados y 50 profesores dio los siguientes resultados:

	Estudiantes	**Empleados**	**Profesores**
Favorable	63	27	30
Desfavorable	37	23	20

Si se desea saber si hay pruebas de una diferencia en la actitud hacia el calendario entre los diversos grupos:

a) Indique cuál es la prueba adecuada para la realización de este contraste, y especifique las hipótesis a contrastar

b) Realice el contraste correspondiente, con un nivel de significación del 1 %.

25. Un estudio sobre tabaquismo en tres ciudades, mediante tres muestras aleatorias de tamaño 100, proporcionó los siguientes resultados:

	Ciudad 1	Ciudad 2	Ciudad 3
Fumadores	13	47	18
No fumadores	87	53	82

¿Se pueden considerar homogéneas las tres poblaciones respecto a sus hábitos fumadores, al nivel 0.05?

Esta parte del capítulo la dedicaremos al análisis de las pruebas no paramétricas que involucran una o dos muestras y cuyo objetivo es probar si los parámetros de los que provienen las muestras adquieren ciertos valores particulares. El lector notará la similitud con las pruebas desarrolladas en el capítulo anterior.

9.4. Pruebas sobre una sola muestra

En las pruebas no paramétricas sobre una muestra se distinguen dos clases: aquellas que contrastan el valor de una medida estadística (de localización, de dispersión, etc.) y las que contrastan una característica general de los datos (ajuste a una ley, aleatoriedad, etc.).

Existe una amplia variedad de estas pruebas, nosotros solo examinaremos los contrastes más comunmente utilizados y que generalmente se presentan en los programas estadísticos.

9.4.1. Prueba de los signos

El contraste de los signos es la prueba no paramétrica más antigua y en ella se basan muchas otras. Se utiliza para contrastar hipótesis sobre el parámetro de localización y en el análisis de comparación de datos pareados. Consideremos una muestra aleatoria de tamaño n tal que sus observaciones estén o puedan estar clasificadas en dos categorías: 0 y 1, + y −, etc.

Podemos establecer hipótesis acerca de la mediana: sabemos que la mediana deja por encima de sí el mismo número de valores que por debajo. Considerando que $x_i - Med > 0$, darán signos positivos (+) y $x_i - Med < 0$ signos negativos (−), en la población original tendremos tantos (+) como (−). Se tratará de ver hasta que punto el número de signos (+) esta dentro de lo que cabe esperar que ocurra por azar si el valor propuesto como mediana es verdadero[1].

Teniendo en cuenta que se trabaja con dos clases de valores, los que están por encima y los que están por debajo de la mediana, los estadísticos de contraste siguen una distribución binomial $\mathbf{Bin}(n, 0.5)$, si se supone independencia y constancia de la probabilidad en el muestreo, ya que la probabilidad de que un valor se encuentre por encima (o por debajo) de la mediana es $p = 0.5$.

[1] Lo mismo se puede decir respecto a los cuartiles, quintiles o deciles.

Si X es la variable aleatoria que cuenta el número de ocurrencias del signo menos frecuente; entonces, su probabilidad se calcula por

$$\Pr(X=k)=\mathbf{C}_n^k p^k(1-p)^{n-k}, \quad p=\frac{1}{2}, \quad k=0,1,\ldots,n.$$

Como nos interesa la ocurrencia de valores tan extremos o más extremos que el observado, la probabilidad deseada es $\Pr(X\le k)$.

Observaciones

1. Si al determinar los signos de las diferencias, obtenemos un valor cero, a éste no se lo considerará el momento de contabilizar el número de signos.
2. Si $n<30$, se utiliza la ley binomial; en cambio, si $n\ge 30$, se utiliza la aproximación mediante la ley normal $Z=\dfrac{X-\frac{n}{2}}{\frac{\sqrt{n}}{2}}\sim\mathcal{N}(0,1)$.
3. Si se desea realizar un contraste para un percentil de orden $100q\,\%$, distinto de la mediana, se cambiará el valor de $p=\frac{1}{2}$ por el valor q correspondiente.

A continuación se exponen las pruebas bilateral y unilateral sobre la mediana.

a) **Prueba bilateral para la mediana.**

1. *Hipótesis Nula.* H_0: $Q_2=\mu_0$.
2. *Hipótesis Alternativa.* H_1: $Q_2\ne\mu_0$.
3. *Estadístico de Prueba.* $p_{obs}=\Pr(X\le k)=\sum_{r=0}^{k}\mathbf{C}_n^r p^r(1-p)^{n-r}$ donde $k<\frac{n}{2}$.
4. *Criterio de Rechazo.* $2p_{obs}<\alpha$.

Ejemplo. En una prueba de aptitud tomada a 12 aspirantes a un puesto en una empresa se obtuvieron los siguientes puntajes:

6.6 6.8 4.4 7.3 8.5 4.5 6.7 6.0 3.4 9.1 5.3 4.8.

Determine si la mediana de las notas es 5.

Solución: Tenemos que $\mu_0=5$; entonces, si restamos a cada dato el valor de prueba, tenemos

1.6 1.8 −0.6 2.3 3.5 −0.5 1.7 1.0 −1.6 4.1 0.3 −0.2.

La secuencia de signos que se obtiene es + + − + + − + + − + + −. El signo menos frecuente es (−), que aparece 4 veces; de manera que la prueba es

1. *Hipótesis Nula.* H_0: $Q_2=5$.
2. *Hipótesis Alternativa.* H_1: $Q_2\ne 5$.
3. *Estadístico de Prueba.* $p_{obs}=\Pr(X\le 4)=\sum_{r=0}^{4}\mathbf{C}_{12}^r\left(\frac{1}{2}\right)^r\left(\frac{1}{2}\right)^{n-r}=0.194.$

4. *Criterio de Rechazo.* $2p_{obs} < 0.05$.
5. *Decisión.* Como $2 \times 0.194 = 0.388 > 0.05$, no se rechaza la hipótesis nula; por lo que se puede asumir que la mediana de las calificaciones es igual a 5.

b) **Prueba unilateral para la mediana.**

1. *Hipótesis Nula.* H_0: $Q_2 = \mu_0$.
2. *Hipótesis Alternativa.* H_1: $Q_2 > \mu_0$ o bien H_1: $Q_2 < \mu_0$.
3. *Estadístico de Prueba.* $p_{obs} = \Pr(X \leq k) = \sum_{r=0}^{k} C_n^r p^r (1-p)^{n-r}$ donde $k < \frac{n}{2}$.
4. *Criterio de Rechazo.* $p_{obs} < \alpha$.

Ejemplo. En un campeonato de fútbol, en que participan muchos equipos se escogió una muestra de 11 equipos. El número de puntos que acumularon estos equipos es el siguiente:

2 4 8 9 12 14 17 21 22 24 26.

Al nivel de significación del 10 %, probar que la mediana de los puntos acumulados por los equipos en el campeonato es menor a 22.

Solución: Tenemos que $\mu_0 = 22$; entonces hay 8 signos negativos, 2 positivos y un cero. Por tanto, $n = 10$ y $k = 2$; de manera que la prueba queda así:

1. *Hipótesis Nula.* H_0: $Q_2 = 22$.
2. *Hipótesis Alternativa.* H_1: $Q_2 < 22$.
3. *Estadístico de Prueba.* $p_{obs} = \Pr(X \leq 2) = \sum_{r=0}^{2} C_{10}^r p^r (1-p)^{10-r} = 0.0547$.
4. *Criterio de Rechazo.* $p_{obs} < 0.1$.
5. *Decisión.* Como $0.0547 < 0.1$, se rechaza H_0. La mediana es menor a 22.

9.4.2. Prueba de los rangos con signos de Wilcoxon

Esta prueba, también conocida como *contraste T de Wilcoxon*, se utiliza para comprobar que la mediana es igual a un valor dado y para su aplicación es necesario que los datos vengan dados en escala ordinal o de intervalo. El procedimiento es el siguiente:

1. Se determinan las diferencias entre cada uno de los valores observados y el valor hipotético de la mediana: $d_i = x_i - \mu_0$.

2. Si alguna de las diferencias es igual a cero, se elimina la observación correspondiente. De esta manera se reduce el tamaño efectivo de la muestra a n, el número de diferencias no nulas.

3. Se ordenan los valores absolutos de las diferencias, de menor a mayor, asignando el rango 1 a la menor diferencia absoluta, 2 a la siguiente diferencia menor, y así sucesivamente. Cuando las diferencias son iguales, se asigna el rango promedio a los valores que son iguales.

4. Se obtienen, por separado, la suma de los rangos para las diferencias negativas (T^-) y positivas (T^+).

5. Si $n \leq 20$ se utiliza la tabla de puntos porcentuales de la prueba de Wilcoxon (Tabla 6); si $n > 20$ se aprovecha que $Z = \dfrac{T - \dfrac{n(n+1)}{4}}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}} \sim \mathcal{N}(0,1)$.

Entonces, el contraste queda así:

a) **Prueba bilateral para la mediana.**

1. *Hipótesis Nula.* H_0: $Q_2 = \mu_0$.
2. *Hipótesis Alternativa.* H_1: $Q_2 \neq \mu_0$.
3. *Estadístico de Prueba.* $T_{obs} = \text{mín}\{T^-, T^+\}$.
4. *Región de Rechazo.* $T_{obs} < T_{\alpha/2}(n)$.

Ejemplo. En una prueba de aptitud tomada a 12 aspirantes a un puesto en una empresa se obtuvieron los siguientes puntajes:

6.6 6.8 4.4 7.3 8.5 4.5 6.7 6.0 3.4 9.1 5.3 4.8.

Determine si la mediana de las notas es 5, para $\alpha = 5\,\%$.

Solución: Para $\mu_0 = 5$, la asignación de rangos se resume en la siguiente tabla:

Valores	$d_i = x_i - \mu_0$	Rango +	Rango −
6.6	+1.6	6.5	
6.8	+1.8	9	
4.4	−0.6		4
7.3	+2.3	10	
8.5	+3.5	11	
4.5	−0.5		3
6.7	+1.7	8	
6.0	+1.0	5	
3.4	−1.6		6.5
9.1	+4.1	12	
5.3	+0.3	2	
4.8	−0.2		1
Total		63.5	14.5

La prueba es

1. *Hipótesis Nula.* H_0: $Q_2 = 5$.
2. *Hipótesis Alternativa.* H_1: $Q_2 \neq 5$.
3. *Estadístico de Prueba.* $T_{obs} = \text{mín}\{14.5, 63.5\} = 14.5$.
4. *Región de Rechazo.* En la tabla de los puntos porcentuales de este contraste se observa que $T_{0.025}(12) = 14$. La región de rechazo es $T_{obs} < 14$.
5. *Decisión.* Se cumple que $14 > 14.5$; por lo tanto, no se rechaza H_0. La mediana de las calificaciones es igual a 5.

b) **Prueba unilateral para la mediana.**

1. *Hipótesis Nula.* H_0: $Q_2 = \mu_0$.
2. *Hipótesis Alternativa.* H_1: $Q_2 > \mu_0$ o bien H_1: $Q_2 < \mu_0$.
3. *Estadístico de Prueba.* $T_{obs} = \text{mín}\{T^-, T^+\}$.
4. *Región de Rechazo.* $T_{obs} < T_\alpha(n)$.

Ejemplo. En un campeonato de fútbol los 11 equipos participantes acumularon el siguiente número de puntos:

$$2, \quad 4, \quad 8, \quad 9, \quad 12, \quad 14, \quad 17, \quad 21, \quad 22, \quad 24, \quad 26.$$

Al nivel de significación del 10 %, probar que la mediana de los puntos acumulados es menor a 22.

Solución: Pongamos $\mu_0 = 22$ y calculemos los valores de T^- y de T^+:

Valores	$d_i = x_i - \mu_0$	**Rango −**	**Rango +**
2	−20	10	
4	−18	9	
8	−14	8	
9	−13	7	
12	−10	6	
14	−8	5	
17	−5	4	
21	−1	1	
22	0	—	—
24	2		2
26	4		3
Total		50	5

Con esto elementos, la prueba es:

1. *Hipótesis Nula.* H_0: $Q_2 = 22$.
2. *Hipótesis Alternativa.* H_1: $Q_2 < 22$.
3. *Estadístico de Prueba.* $T_{obs} = \text{mín}\{50, 5\} = 5$.
4. *Región de Rechazo.* Aquí, $T_{0.1}(10) = 15$, la región es $T_{obs} < 15$.
5. *Decisión.* Se rechaza H_0 y deducimos que la mediana es menor a 22.

9.4.3. Prueba de Kolmogorov-Smirnov de ajuste a una ley de probabilidad

El método de Kolmogorov-Smirnov es un procedimiento utilizado para comprobar la hipótesis nula de que la muestra procede de una población que está distribuida según una ley de probabilidad específica.

El estadístico de prueba, que se denota por D_{obs}, se define por

$$D_{obs} = \text{máx}\,|F_0(x) - S_n(x)|\,,$$

donde $F_0(x)$ y $S_n(x)$ son las probabilidades acumuladas esperadas y observadas, respectivamente.

La prueba estadística es la siguiente:

1. *Hipótesis Nula.* H_0: La población sigue una $\mathcal{L}(p)$ dada.

2. *Hipótesis Alternativa.* H_1: La población no sigue una ley $\mathcal{L}(p)$ dada.

3. *Estadístico de Prueba.* $D_{obs} = \text{máx}\,|F_0(x) - S_n(x)|$.

4. *Región de Rechazo.* $D_{obs} > D_\alpha(n)$.

Los valores de $D_\alpha(n)$ se encuentran tabulados para diversos valores de α y n. (Tabla 7)

Ejemplos

1. Tras jugar a los dados, un apostador comenzó a sospechar que el casino hacía trampa. Decidió, por tanto, anotar las tiradas del dado con que jugaba, obteniendo la siguiente tabla:

Número	1	2	3	4	5	6
Frecuencia	16	10	16	11	32	15

¿Se puede afirmar que el dado es incorrecto?

Solución: Formemos la tabla de frecuencias relativas acumuladas, adjuntando las columnas de la probabilidad teórica y de la diferencia:

x_i	**Frecuencia absoluta** (n_i)	**Frecuencia relativa** (f_i)	**Frec. relativa acumulada** ($S_n(x_i)$)	$F_0(x_i)$	$\|F_0(x_i) - S_n(x_i)\|$
1	16	0.16	0.16	0.1667	0.0067
2	10	0.10	0.26	0.3333	0.0733
3	16	0.16	0.42	0.5000	0.0800
4	11	0.11	0.53	0.6667	0.1367
5	32	0.32	0.85	0.8333	0.0167
6	15	0.15	1.00	1.000	0

La prueba de ajuste queda de la siguiente manera:

1. *Hipótesis Nula.* H_0: Los datos siguen una ley uniforme discreta con $p = 1/6$.
2. *Hipótesis Alternativa.* H_1: Los datos no siguen una ley uniforme discreta con $p = 1/6$.
3. *Estadístico de Prueba.* $D_{obs} = \text{máx}\,|F_0(x) - S_n(x)| = 0.1367$.
4. *Región de Rechazo.* En la tabla del contraste K-S encontramos que $D_{0.05}(100) = 0.136$. Se define la región de rechazo $D_{obs} > 0.136$.
5. *Decisión.* Como $D_{obs} > D_\alpha(n)$, se rechaza H_0 y concluimos que el dado está cargado.

2. En una empresa, el salario mensual de los trabajadores se distribuye según la siguiente tabla:

Desde	**Hasta**	**No. empleados**
150	190	6
190	230	16
230	270	47
270	310	55
310	350	38
350	390	19
390	430	14
430	470	5

Comprobar si se puede asegurar que el salario mensual en dicha empresa puede considerarse una variable $\mathcal{N}(306, 58)$.

Solución: Completemos la tabla de frecuencias relativas acumuladas, calculando las dos distribuciones (la muestral y la teórica), teniendo en cuenta que para calcular $F_0(x)$ deberemos utilizar la tabla de la ley normal.

Desde	Hasta	No. empleados	N_i	$S_n(x_i)$	$F_0(x_i)$	$\lvert F_0(x_i) - S(x_i)\rvert$
150	190	3	3	0.0205	0.0241	0.0036
190	230	13	16	0.1096	0.1089	0.0007
230	270	31	47	0.3219	0.2976	0.0243
270	310	44	91	0.6233	0.5636	0.0597
310	350	30	121	0.8288	0.8001	0.0287
350	390	14	135	0.9247	0.9348	0.0101
390	430	9	144	0.9863	0.9826	0.0037
430	470	2	146	1.000	0.9935	0.0065

De manera que la prueba es:

1. *Hipótesis Nula.* H_0: Los sueldos siguen una ley $\mathcal{N}(306, 58)$.
2. *Hipótesis Alternativa.* H_1: Los sueldos no siguen una ley $\mathcal{N}(306, 58)$.
3. *Estadístico de Prueba.* $D_{obs} = \text{máx}\,|F_0(x) - S_n(x)| = 0.0597$.
4. *Región de Rechazo.* En la tabla encontramos que $D_{0.05}(146) = \dfrac{1.36}{\sqrt{146}} = 0.1126$; entonces, la región es $D_{obs} > 0.1126$.
5. *Decisión.* No se rechaza H_0; consecuentemente, los salarios siguen la ley $\mathcal{N}(306, 58)$.

9.4.4. Prueba de la aleatoriedad de la muestra

En muchos estudios experimentales es necesario conocer si las muestras obtenidas son aleatorias o, por el contrario, si ellas mantienen algún tipo de correlación o si han sido alteradas. La prueba que nos permite realizar esta identificación también se conoce como *constraste de rachas.* Por extensión, esta prueba es indicada para probar la independencia entre observaciones.

Definición (de racha). Una *racha* es una sucesión de valores por encima o por debajo de la mediana.

La longitud de una racha es el número de observaciones consecutivas que tienen esta propiedad.

Por ejemplo, si las observaciones son

$$6.6 \quad 6.8 \quad 4.4 \quad 7.3 \quad 8.5 \quad 4.5 \quad 6.7 \quad 6.0 \quad 3.4 \quad 9.1 \quad 5.3 \quad 4.8,$$

la mediana es $Q_2 = 6.3$, luego de representar por $(-)$ los valores inferiores a la mediana y por $(+)$ a los superiores, obtenemos la secuencia: $+, +, -, +, +, -, +, -, -, +, -, -$. Existen 8 rachas de longitudes 2, 1, 2, 1, 1, 2, 1, 2.

Sean r_{obs} el número de rachas en una sucesión ordenada de observaciones; m y n el número de observaciones que están por debajo y por encima de la mediana de los datos, respectivamente. Con estos elementos el contraste queda de la siguiente manera:

1. *Hipótesis Nula.* H_0: Las observaciones aparecen de manera aleatoria.
2. *Hipótesis Alternativa.* H_1: Las observaciones no aparecen de manera aleatoria.

3. *Estadístico de Prueba.* r_{obs}.
4. *Región de Rechazo.* $r_{obs} < l_\alpha(m, n)$ o $r_{obs} > u_\alpha(m, n)$.

Los valores de comparación para esta prueba se encuentran tabulados (Tabla 5).

Ejemplo. En una prueba de aptitud tomada a 12 aspirantes a un puesto en una empresa se obtuvieron los siguientes puntajes:

$$6.6 \quad 6.8 \quad 4.4 \quad 7.3 \quad 8.5 \quad 4.5 \quad 6.7 \quad 6.0 \quad 3.4 \quad 9.1 \quad 5.3 \quad 4.8.$$

Determine si los datos aparecen de manera aleatoria, para $\alpha = 5\,\%$.

Solución: La mediana de estos datos es 6.3. Si a cada uno de los datos le restamos la mediana, queda la siguiente secuencia de signos: $+, +, -, +, +, -, +, -, -, +, -, -$.

Entonces, el número de rachas es $r_{obs} = 8$ y $m = n = 6$.

1. *Hipótesis Nula.* H_0: Las observaciones aparecen de manera aleatoria.
2. *Hipótesis Alternativa.* H_1: Las observaciones no aparecen de manera aleatoria.
3. *Estadístico de Prueba.* $r_{obs} = 8$.
4. *Región de Rechazo.* En la tabla, vemos los valores críticos con $\alpha = 5\,\%$: $3 \leq r_{obs} \leq 10$.
5. *Decisión.* Como $r_{obs} = 8$ cae en la región de aceptación, podemos afirmar que los valores aparecen aleatoriamente.

9.4.5. Prueba para identificar valores atípicos

Esta prueba permite identificar la presencia de valores extremos en un conjunto de datos. Para la detección de valores atípicos, en general, se manejan criterios empíricos; por ejemplo, que una observación esté alejada una distancia de más de 3 desviaciones estándar de la media, pero estos criterios no tienen fundamento técnico y descuidan la influencia del tamaño de la muestra.

Este contraste detecta un valor atípico a la vez y cada uno debe retirarse de la muestra iterativamente hasta que no se detecten más valores atípicos. La prueba fue desarrollada por Grubbs[2] y se basa en la suposición que los datos siguen de una ley normal o que pueden ser aproximados razonablemente por ésta[3].

La prueba de hipótesis es la siguiente:

1. *Hipótesis Nula.* H_0: No hay valores atípicos en el conjunto de datos.
2. *Hipótesis Alternativa.* H_1: Hay al menos un valor atípico en el conjunto de datos.
3. *Estadístico de Prueba.* $g_{obs} = \dfrac{\max\limits_i\{|x_i - \overline{x}|\}}{s}$.
4. *Región de Rechazo.* $g_{obs} > G_\alpha(n)$.

[2] Grubbs, F. E. (1969), "Procedures for Detecting Outlying Observations in Samples," *Technometrics*, 11, 1-21.

[3] Aunque éste es un contraste paramétrico, se lo incluye en este capítulo porque no se realiza sobre el valor de un parámetro sino sobre una característica general de los datos.

La tabla de puntos porcentuales para la prueba de Grubbs se encuentra en el Apéndice (Tabla 8).

Ejemplo. En una prueba de aptitud tomada a 12 aspirantes a un puesto en una empresa se obtuvieron los siguientes puntajes:

6.6 6.8 4.4 7.3 8.5 4.5 6.7 6.0 3.4 9.1 5.3 4.8.

Determine si existe algún dato atípico. Use $\alpha = 5\,\%$.

Solución: El promedio y la desviación estándar son $\overline{x} = 6.117$ y $s = 1.716$ y el dato que está más alejado del promedio es 9.1.

Con estos elementos, la prueba queda así:

1. *Hipótesis Nula.* H_0: No hay valores atípicos en el conjunto de datos.
2. *Hipótesis Alternativa.* H_1: Hay al menos un valor atípico en el conjunto de datos.
3. *Estadístico de Prueba.* $g_{obs} = \dfrac{\max_i\{|x_i - \overline{x}|\}}{s} = \dfrac{|9.1 - 6.117|}{1.716} = 1.74.$
4. *Región de Rechazo.* En la tabla de la puntos porcentuales se lee que $G_{0.05}(12) = 2.41$ y la región de rechazo es $g_{obs} > 2.41$.
5. *Decisión.* Como $1.74 < 2.29$, entonces $x = 9.1$ no es un valor atípico; es decir, la muestra no contiene valores atípicos.

9.5. Pruebas sobre dos muestras

Las pruebas no paramétricas sobre dos muestras que examinaremos son análogas a las paramétricas para comparar dos medias; es decir, para datos emparejados y para muestras independientes. También, se presentará una versión no paramétrica del coeficiente de correlación.

9.5.1. Prueba de los signos para datos emparejados

Esta prueba se aplica cuando se tienen dos muestras relacionadas y se quiere probar la hipótesis de que las dos poblaciones tienen medianas iguales, pero no se quieren realizar suposiciones acerca de la normalidad de los datos. Los datos deben presentarse en al menos escala ordinal.

Sea (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) una muestra aleatoria de pares de observaciones; donde (x_i, y_i) representa dos mediciones tomadas de la misma unidad muestral, antes y después de un tratamiento o fenómeno que la afectó. Se construye una muestra de las diferencias d_1, d_2, ..., d_n, donde $d_i = x_i - y_i$ $(i = 1, 2, \ldots, n)$ y se aplica la prueba del signo para una muestra; es decir, contamos el número de observaciones negativas y positivas y determinamos cuál es el signo que menos aparece y cuántas veces lo hace (k). Además, llamaremos Q_x a la mediana de la población X y Q_y a la mediana de la población Y. Con ésto, las pruebas son las siguientes:

a) **Prueba bilateral para la diferencia de medianas.**

1. *Hipótesis Nula.* H_0: $Q_x = Q_y$.
2. *Hipótesis Alternativa.* H_1: $Q_x \neq Q_y$.

3. *Estadístico de Prueba.* $p_{obs} = \Pr(X \leq k) = \sum_{r=0}^{k} \mathbf{C}_n^r p^r (1-p)^{n-r}$, donde $k < \dfrac{n}{2}$.
4. *Criterio de Rechazo.* $2p_{obs} < \alpha$.

b) **Prueba unilateral para la diferencia de medianas.**

1. *Hipótesis Nula.* H_0: $Q_x = Q_y$.
2. *Hipótesis Alternativa.* H_1: $Q_x > Q_y$ o bien H_1: $Q_x < Q_y$.
3. *Estadístico de Prueba.* $p_{obs} = \Pr(X \leq k) = \sum_{r=0}^{k} \mathbf{C}_n^r p^r (1-p)^{n-r}$, donde $k < \dfrac{n}{2}$.
4. *Criterio de Rechazo.* $p_{obs} < \alpha$.

Ejemplo. En un estudio para determinar la aceptación de un tipo de yogurt, se pidió a 8 voluntarios que calificaran a dos versiones (antigua y nueva) del producto. Las calificaciones y sus diferencias se resumen en la tabla.

Individuo	Versión antigua	Versión nueva	Diferencia	Signo de la diferencia
1	6	8	-2	$-$
2	4	9	-5	$-$
3	5	4	$+1$	$+$
4	8	7	$+1$	$+$
5	3	9	-6	$-$
6	6	9	-3	$-$
7	7	7	0	0
8	5	9	-4	$-$

Verificar si hubo incremento en la calificación al presentar la nueva versión.

Solución: Tenemos que $n = 7$ y $k = 2$. Entonces, la prueba queda así:

1. *Hipótesis Nula.* H_0: $Q_x = Q_y$.
2. *Hipótesis Alternativa.* H_1: $Q_x < Q_y$.
3. *Estadístico de Prueba.* $p_{obs} = \Pr(X \leq 2) = \sum_{r=0}^{k} \mathbf{C}_7^r p^r (1-p)^{7-r} = 0.2266$.
4. *Criterio de Rechazo.* $p_{obs} < 0.05$.
5. *Decisión.* Se tiene que $p_{obs} = 0.0226 > 0.05$ y se acepta H_0. No hubo cambio significativo en la apreciación de las dos versiones del producto.

Observación. También se puede aplicar la prueba de Wilcoxon a datos emparejados, mediante la aplicación de este contraste a la diferencia entre los valores de las dos muestras. Se recomienda que el lector formule esta prueba.

9.5.2. Prueba de Mann-Whitney para dos muestras independientes

La prueba U de Mann-Whitney se emplea para probar que dos muestras independientes provienen de la misma población o de poblaciones que siguen la misma ley de probabilidad.

Supongamos que se dispone de dos muestras independientes, X y Y, cuyas funciones de distribución son $F_X(x)$ y $F_Y(x)$, respectivamente. Diremos, también, que el tamaño de la muestra X es m y que el tamaño de la muestra Y es n. Entonces, para realizar el contraste se realizan los siguientes pasos:

1. Se combinan las dos muestras en una sola.
2. Se asignan rangos a la muestra combinada de las dos muestras. Si se producen empates, se asignará el promedio de los rangos a las observaciones empatadas.
3. Se suman los rangos de las dos muestras y se calculan los estadísticos

$$U_1 = mn + \frac{m(m+1)}{2} - \sum \text{Rangos}(X), \quad U_2 = mn + \frac{n(n+1)}{2} - \sum \text{Rangos}(Y).$$

Con estos elementos, se realiza la siguiente prueba:

1. *Hipótesis Nula.* H_0: $F_X(x) = F_Y(x)$
2. *Hipótesis Alternativa.* H_1: $F_X(x) \neq F_Y(x)$
3. *Estadístico de Prueba.* $U_{obs} = \text{mín}\{U_1, U_2\}$.
4. *Región de Rechazo.* $U_{obs} \leq U_\alpha(m, n)$.

Los valores de comparación para esta prueba se encuentran tabulados (Tabla 10).

Ejemplo. El dueño de un almacén registró las ventas semanales de sus dos empleados y quiere saber si ellos pueden considerarse iguales como vendedores. En la siguiente tabla se numeran las ventas de cada dependiente.

Empleado A	197	194	188	185	180	173	169	169
Empleado B	190	166	175	172	167	180	160	

Solución: Los rangos asignados a las muestras son:

Empleado A	Empleado B	Rangos A	Rangos B
197	190	15	13
194	166	14	2
188	175	12	8
185	172	11	6
180	167	9.5	3
173	180	7	9.5
169	160	4.5	1
169		4.5	
$m = 8$	$n = 7$	$\sum$ Rangos = 77.5	$\sum$ Rangos = 42.5

Los estadísticos U_1 y U_2 son

$$\begin{aligned} U_1 &= mn + \frac{m(m+1)}{2} - \sum \text{Rangos}(A) = 8 \times 7 + \frac{8 \times 9}{2} - 77.5 = 14.5, \\ U_2 &= mn + \frac{n(n+1)}{2} - \sum \text{Rangos}(B) = 8 \times 7 + \frac{7 \times 8}{2} - 42.5 = 41.5. \end{aligned}$$

De manera que la prueba queda así:

1. *Hipótesis Nula.* H_0: $\mu_1 = \mu_2$
2. *Hipótesis Alternativa.* H_1: $\mu_1 \neq \mu_2$
3. *Estadístico de Prueba.* $U_{obs} = \text{mín}\{U_1, U_2\} = 14.5$.
4. *Región de Rechazo.* Para $\alpha = 0.01$, $U_{0.01}(8, 7) = 8$; entonces, la región es $U_{obs} > 8$.
5. *Decisión.* Como $14.5 > 8$, se rechaza H_0; es decir, los dos empleados venden iguales cantidades.

9.5.3. Prueba de correlación de rangos de Spearman

El *coeficiente de correlación de rangos ordenados de Spearman* fue el primer estadístico basado en rangos que se desarrolló. Se lo emplea para determinar la existencia de asociación entre dos variables de tipo ordinal.

Si disponemos de n parejas de observaciones (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) de dos variables X y Y y asignamos rangos a la primera (R_{x_i}) y a la segunda componente (R_{y_i}) de las parejas, se define el coeficiente de correlación de Spearman como

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)},$$

donde $d_i = R_{x_i} - R_{y_i}$ es la diferencia entre los rangos asignados a x_i y y_i.

El procedimiento para su cálculo es el siguiente:

1. Se asignan rangos a las observaciones de las variables X y Y.
2. Si hubiera empates, se asigna el promedio de los rangos a los individuos empatados.
3. Se determina el valor de d_i para cada individuo y aplique la fórmula de r_s.

El coeficiente de correlación de Spearman tiene una interpretación similar al de Pearson:

- Toma valores entre -1 y $+1$.
- Cuando el valor absoluto de este coeficiente es cercano a 1 indica que hay asociación entre las dos variables.
- Cuando el valor del coeficiente es cercano a cero, indica que hay poca o ninguna asociación entre las variables.

La ventaja de este coeficiente es que no requiere realizar suposiciones de normalidad.

El contraste correspondiente al coeficiente de Spearman es:

1. *Hipótesis Nula.* H_0: $\rho = 0$
2. *Hipótesis Alternativa.* H_1: $\rho \neq 0$
3. *Estadístico de Prueba.* r_s
4. *Región de Rechazo.* $r_s > r_\alpha(n)$

La tabla para realizar este contraste se encuentra en la Tabla 9 del Apéndice.

Ejemplo. Se examinó el registro de notas de 10 alumnos en las materias de Álgebra y Cálculo. Las calificaciones correspondientes a cada uno se da en la siguiente tabla.

Alumno	1	2	3	4	5	6	7	8	9	10
Álgebra	17	18	19	12	23	23	25	26	31	33
Cálculo	21	14	27	18	20	25	34	32	39	33

Verificar si existe asociación entre las notas en las dos materias.

Solución: Formemos la tabla de rangos y su diferencia:

Alumno	Álgebra	Cálculo	Rangos A (R_{x_i})	Rangos C (R_{y_i})	Diferencia d_i
1	17	21	2	4	−2
2	18	14	3	1	2
3	19	27	4	6	−2
4	12	18	1	2	−1
5	23	20	5.5	3	2.5
6	23	25	5.5	5	0.5
7	25	34	7	9	−2
8	26	32	8	7	1
9	31	39	9	10	−1
10	33	33	10	8	2
					$\sum_{i=1}^{n} d_i^2 = 29.5$

Entonces, el coeficiente de correlación es

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\,(n^2-1)} = 1 - \frac{6 \times 29.5}{10(100-1)} = 0.821.$$

La prueba de hipótesis de nulidad de la correlación es:

1. *Hipótesis Nula.* H_0: $\rho = 0$
2. *Hipótesis Alternativa.* H_1: $\rho \neq 0$
3. *Estadístico de Prueba.* $r_s = 0.821$.
4. *Región de Rechazo.* En la tabla del contraste tenemos que $r_{0.05}(10) = 0.648$. La región de rechazo es $r_s > 0.648$.
5. *Decisión.* Se tiene que $0.821 > 0.648$ y se rechaza H_0. Entonces, hay asociación entre las notas de las dos materias.

9.6. Ejercicios

Pruebas sobre una sola muestra

1. En una comisaría se registraron el número de denuncias diarias por robos:

   17 15 20 20 25 32 28 25 35 12 26 24

   Utilice un nivel de significación del 5 % para probar que la mediana del número de denuncias no es mayor a 20 por día:

   a) mediante la prueba de los signos;

b) mediante la prueba de Wicoxon.

2. En un estudio sobre las remesas enviadas por los parientes emigrantes a varias familias dio los siguientes resultados (en miles de dólares):

2.4 2.3 1.7 1.2 2.5 3.6 4.2 2.2 2.3 3.1

Con el empleo de $\alpha = 0.05$, pruebe si la mediana de estas remesas es diferente de 2700 dólares:

a) mediante la prueba de los signos;

b) mediante la prueba de Wicoxon.

3. Para ingresar a una agencia bancaria se formó una cola formada por H (hombres) y M (mujeres), que está formada así:

H M H M H H H M H H H M M H M M H H H M

Determine la aleatoriedad en la conformación de la cola.

4. Un profesor registró el número de estudiantes ausentes a sus clases durante 24 días consecutivos.

38	31	32	27	28	30	26	33	36	30	28	35
32	29	35	31	33	31	28	30	28	25	29	33

a) Estudie la aleatoriedad de la muestra con un nivel de significación de 0.05;

b) Se puede asegurar que la mediana es igual a 27.

5. Un corredor de bolsa observó la siguiente venta de bonos a lo largo de un año:

Enero	19	Julio	22
Febrero	23	Agosto	24
Marzo	20	Septiembre	25
Abril	17	Octubre	28
Mayo	18	Noviembre	30
Junio	20	Diciembre	21

a) Aplique una prueba de rachas para decidir si los datos pueden considerarse aleatorios;

b) ¿Es posible afirmar que la mediana es igual a 23?

6. Se desea ajustar la distribución de la cantidad de correos electrónicos que recibe el servidor de una empresa durante un intervalo de cinco minutos. Para 400 intervalos de cinco minutos se contaron la cantidad de mensajes recibidos y se obtuvieron las siguientes frecuencias:

No. correos	0	1	2	3	4	5	6	7	8	9	10
Frecuencia	7	11	47	76	68	74	46	39	15	9	8

¿Los datos pueden considerarse distribuidos con la distribución de Poisson con $\lambda = 4.6$?

7. Los siguientes datos corresponden a los tiempos de duración (medidos en meses) de lámparas marca Lucky:

3.33 6.71 2.53 7.15 16.82 3.56 0.17 2.15 3.51 7.24.

a) Estudie si es razonable suponer que los datos corresponden a una distribución exponencial;

b) ¿Se puede suponer que la mediana de la muestra es igual a cuatro?

8. Los siguientes datos corresponden a los tiempos de vida (en semanas) de colonias de bacterias criadas en un laboratorio bajo condiciones estables de temperatura y humedad:

3.31 5.48 1.37 7.44 2.73 4.19 3.18 2.23.

En todas las pruebas use el nivel de significación del 10 %:

a) Estudie la aleatoriedad de la muestra;
b) ¿Es razonable suponer que los datos corresponden a una distribución exponencial?;
c) ¿Existe algún valor atípico en la muestra?

9. Se registró el caudal promedio (en m^3/s) que ingresó a una central hidroeléctrica durante 12 días consecutivos:

99.0 79.8 77.4 87.8 59.0 71.0 51.4 92.1 69.4 68.9 70.1 90.2.

a) ¿Es razonable suponer que los datos son aleatorios?;
b) ¿Es razonable suponer que los datos tienen distribución normal?;
c) Mediante la prueba de Wilcoxon, determine si la media de la distribución es $75\,\mathrm{m}^3/\mathrm{s}$;
d) Determine si existe algún valor atípico.

10. Se tiene la función densidad dada por $f(x) = \begin{cases} 1-|x|, & \text{si } -1 \leq x \leq 1; \\ 0, & \text{caso contrario.} \end{cases}$

Se dispone de una muestra de mediciones de esta variable aleatoria:

0.03 0.01 0.32 0.88 –0.41 –0.18.

Utilizando las herramientas estadísticas adecuadas responda a las siguientes preguntas:

a) ¿Es razonable suponer que la muestra es aleatoria?;
b) ¿Es razonable suponer que la distribución de la muestra corresponde a la densidad dada anteriormente?;
c) Independientemente del resultado de la parte anterior, ¿se puede suponer que los datos provienen de una distribución centrada en 0?

11. Se dispone de una máquina para llenar latas con $10\,\mathrm{cm}^3$ de aceite (las latas llenadas al borde tienen una capacidad de $11\,\mathrm{cm}^3$). Un operario afirma haberla ajustado de modo que la cantidad de aceite que envía el pico de la máquina envasadora es de $(10+\varepsilon)\,\mathrm{cm}^3$, donde ε es una variable aleatoria con la distribución $\mathcal{U}\left(-\frac{1}{10}, \frac{1}{10}\right)$. Para verificar la afirmación anterior se estudió el contenido de 12 latas obteniendo los siguientes valores de ε:

0.098 0.068 -0.004 –0.026 –0.012 0.096 0.050 0.062 0.036 0.04 0 0.054.

a) Determine si la muestra es aleatoria;
b) ¿Es razonable suponer que la distribución de la muestra es $\mathcal{U}\left(-\frac{1}{10}, \frac{1}{10}\right)$?;
c) Si la distribución del error fuese $\mathcal{U}\left(-\frac{1}{10}, \frac{1}{10}\right)$ el error medio sería $0\,\mathrm{cm}^3$. ¿Los datos experimentales confirman la afirmación anterior?

12. Un fabricante de fusibles asegura que, con una sobrecarga del 20 %, los tiempos de vida de sus fusibles (desde que se conectan con sobrecarga hasta que se funden) se distribuyen uniformemente entre 10 minutos y 15 minutos. Para probar esta afirmación una muestra de 8 fusibles fue sometida a una sobrecarga del 20 %. Los tiempos en que tardaron en fundirse dichos fusibles fueron los siguientes:

13.34 10.69 13.37 11.14 13.87 13.75 10.76 12.63.

a) Analice la aleatoriedad de la muestra;

b) ¿Es razonable suponer que la distribución de la muestra es $\mathcal{U}[10, 15]$?;

c) Si la distribución de los tiempos de vida fuese $\mathcal{U}[10, 15]$, el tiempo medio de vida sería de 12.5 min. ¿Los datos experimentales confirman la afirmación anterior?

13. En una empresa operadora de tarjetas de crédito se está analizando el historial del último año del consumo mensual de un cliente. Se conoce que el promedio de sus consumos es 645 dólares y desviación estándar de 148 dólares. Si se encontró que este mes tuvo un consumo de 3200 dólares, ¿debe considerarse que el cliente tuvo un comportamiento atípico?

14. La población ecuatoriana adulta tiene una estatura promedio de 162 cm y desviación estándar de 7.5 cm. En una muestra de 25 personas, ¿a partir de qué valores de la estatura podemos considerar que hay presencia de valores atípicos? (Use $\alpha = 0.05$).

Pruebas sobre dos muestras

15. En dos laboratorios se realizaron mediciones del punto de ebullición de un compuesto de silicio (en °C) de 8 muestras diferentes.

Muestra	1	2	3	4	5	6	7	8
Medición 1	99.78	99.17	100.06	100.14	99.43	100.60	100.59	99.98
Medición 2	100.16	100.09	99.91	100.36	99.71	101.09	99.93	100.06

Determine si los dos laboratorios entregan iguales resultados, usando $\alpha = 0.05$.

16. Se presume que un tratamiento reduce el peso de las personas. Mediante una muestra aleatoria se seleccionan 10 personas que siguieron dicho tratamiento durante todo el tiempo exigido. En la siguiente tabla se presenta el peso de cada paciente, antes y después del tratamiento (medido en kg).

Persona	1	2	3	4	5	6	7	8	9	10
Antes	108	72	81	104	69	73	114	86	92	98
Después	95	76	69	81	56	81	92	81	71	97

Verifique si el tratamiento es efectivo. (Use $\alpha = 5\,\%$)

17. Se tienen dos muestras, X y Y, independientes entre sí, que corresponden a mediciones de niveles de contaminación sonora en las ciudades de Quito y Guayaquil. Se desea saber si ambas ciudades tienen similares niveles de contaminación o si alguna de ellas presenta niveles significativamente mayores.

X	70.1	70.4	75.8	67.5	68.4	73.6	76.9	75.7	71.4	70.3	72.1	69.8		
Y	74.3	74.1	75.4	67.4	69.3	70.5	70.1	69.9	68.7	70.3	70.7	71.1	74.4	70.2

18. Para calcular la velocidad de cálculo de dos computadoras A y B, se realizaron en ambas 5 operaciones. Los tiempos invertidos, en milisegundos, fueron:

A	110	125	141	113	182
B	102	120	135	114	175

Analice si hay diferencias entre las localizaciones de las dos muestras. Use $\alpha = 5\,\%$.

19. Los siguientes datos son los tiempos de duración (medidos en meses) de 10 lámparas marca Lucky:

3.33 6.71 2.53 7.15 16.82 3.56 0.17 2.15 3.51 7.24.

Se dispone ahora de una nueva muestra, independiente de la anterior, de los tiempos de duración de lámparas de marca Wizard:

5.60 3.03 1.46 16.52 0.37 0.58 3.38 9.15 0.90 4.44.

Implemente una prueba de comparación de muestras para concluir si es razonable suponer que las nuevas lámparas tienen la misma duración que las anteriores.

20. Los siguientes datos corresponden a los tiempos de vida (medidos en nanosegundos) de partículas radioactivas emitidas por cierto material:

20.05 1.30 2.54 1.95 9.20 4.20 1.84 1.02 5.60 1.80.

Para un nivel de significación del 10 %, estudie:

a) la aleatoriedad de la muestra;

b) si es razonable suponer que los datos corresponden a una ley de Pareto, caracterizada por la siguiente función de distribución:

$$F(x) = \begin{cases} 1 - \dfrac{1}{x}, & \text{si } x > 1; \\ 0, & \text{caso contrario.} \end{cases}$$

c) Se dispone ahora de una nueva muestra independiente de la anterior correspondiente a los tiempos de vida de partículas radioactivas emitidas por otro material:

2.60 4.81 12.79 3.82 1.21 1.68 15.50 12.03 1.70 9.40.

Realice una prueba de comparación de muestras para concluir si es razonable o no suponer que los nuevos datos tienen la misma distribución que los anteriores.

21. Se tienen dos muestras independientes entre sí, correspondientes a los tiempos de duración (en años) de sistemas electrónicos de marcas distintas. Se desea saber si los equipos de ambas marcas tienen niveles similares de duración o si alguna de ellas presenta una duración significativamente mayor.

Muestra X	2.00	0.67	0.58	1.46	0.28	0.43	1.02	0.96		
Muestra Y	0.20	1.73	0.30	0.02	0.05	1.57	1.46	0.14	0.32	0.60

Utilizando un nivel de significación del 10 %:

a) Aplique la prueba de rachas a las muestras para decidir si se pueden suponer aleatorias;

b) Aplique a ambas muestras las pruebas de signos y de rangos signados de Wilcoxon para decidir si los valores $X_m = 0.69$ y $Y_m = 0.35$ son valores aceptables para las medianas;

c) Pruebe si la mediana de la muestra X es mayor que la de la muestra Y.

22. En una investigación de mercado se pidió a dos niños que calificaran a 10 juguetes en orden de preferencia (de 1 a 10), obteniéndose la siguiente tabla:

Juguete	A	B	C	D	E	F	G	H	I	J
Niño 1	8	9	6	2	1	4	5	7	3	10
Niño 2	7	10	8	5	3	2	4	6	1	9

Determine si las calificaciones están correlacionadas.

23. En un estudio sobre comercio internacional, se ordenaron de manera decreciente a los principales socios comerciales sudamericanos de Ecuador y Argentina.

Socio	Col.	Per.	Bra.	Ven.	Chi.	Bol.	Par.
Ecuador	1	2	3	4	5	6	7
Argentina	3	4	1	5	2	7	6

Calcule el coeficiente de correlación de Spearman y verifique si hay asociación entre las ordenaciones.

24. Las siguientes son las calificaciones obtenidas por 11 jóvenes que se sometieron a evaluaciones en destreza manual y memoria visual.

D. manual	1	3	4	5	6	8	10	11	13	14	16
M. visual	11	15	18	16	23	31	39	56	45	43	37

Realice una prueba para determinar si existe correlación entre las calificaciones en las dos evaluaciones.

25. Un grupo de investigadores desea evaluar si un nuevo equipo de tratamiento aguas residuales es efectivo para reducir los niveles de contaminación de las aguas vertidas a un río por 10 curtiembres. A tales efectos se midió el nivel de contaminantes antes y después del tratamiento, los resultados fueron los siguientes:

Planta	1	2	3	4	5	6	7	8	9	10
Antes	1.52	2.92	4.44	4.24	1.72	3.70	3.64	4.82	2.12	2.08
Después	2.08	3.03	0.80	0.96	2.71	2.39	3.07	2.87	0.33	1.76

a) Determine si efectivamente se ha producido una reducción en los niveles de contaminación;

b) ¿Las dos muestras están correlacionadas?

Capítulo 10

Regresión Lineal Simple

Todos los modelos están equivocados, pero algunos son útiles
G. E. P. Box

Muchas de las aplicaciones estadísticas requieren la estimación de las relaciones existentes entre dos o más variables; por ejemplo, puede ser necesario responder a las preguntas ¿cómo varía el precio promedio anual del maíz, según la producción a nivel nacional? o ¿cómo varía el consumo de gasolina de un auto, según su peso y la potencia del motor? El interés se centra, entonces, en determinar una ecuación que relacione una variable dada con una o más variables que contienen información sobre la primera. A estos problemas dedicaremos los dos siguientes capítulos; antes revisemos algo de la historia de esta parte de la estadística.

No se conoce, con exactitud, quién y cuándo empezó a tratar de expresar algebraicamente las relaciones entre dos o más variables, de las cuales solo se dispone de un conjunto de observaciones; pero en los escritos de Leonardo da Vinci, cuando él trata de las proporciones del cuerpo humano, se encuentran expresiones aritméticas que relacionan las medidas de diversas partes del cuerpo.

Un intento, que está bien documentado, data de 1755, cuando Boscovich y Christopher Maire estaban encargados de medir la longitud del arco de meridiano que pasa por Roma. Boscovich concibió un método para encontrar un modelo que relacione los datos correspondientes a dos variables, mediante el empleo de las llamadas «regresiones elementales». Esta técnica fue mejorada por su autor en 1760 y puesta en una forma más estructurada por Laplace, unos años más tarde.

En 1805, Legendre publicó una obra de astronomía, en la que describió el método de los mínimos cuadrados y lo aplicó al ajuste de datos observacionales. También, hay una serie de artículos presentados por C. F. Gauss a la Sociedad Real de Götinga en los que describe el método de mínimos cuadrados. Luego, en 1885, Sir F. Galton presentó en la revista *Nature* el desarrollo completo de esta técnica, aplicada a lo que él denominó *modelos de regresión*. A partir de esta fecha se mejoró y se completó la técnica, haciendo que ella sea la de mayor empleo en el ajuste de conjuntos de datos.

Actualmente, la construcción de modelos lineales es la base de todas las técnicas estadísticas de naturaleza predictiva y su aplicación se ha extendido a prácticamente todas las ciencias.

10.1. Modelos deterministas y probabilísticos

Existen aplicaciones en las que se dispone de un modelo que presenta una relación exacta entre las variables de interés; por ejemplo, la ley que describe el tiempo que tarda en caer un objeto desde una

altura dada, o la fórmula que nos indica el interés ganado por un capital, dados la tasa de interés y el periodo de la inversión. Tales modelos se denominan *deterministas*.

Sin embargo, en la vida diaria, rara vez se presentan fenómenos que reproducen con exactitud una ley, ya sea porque existen errores en la medición o porque hay otras variables que no son consideradas, por su escasa influencia, pero que son suficientes para que el modelo propuesto no sea exacto.

Un modelo en el que una o más variables es de naturaleza aleatoria se denomina *probabilístico* y a la determinación y examen de la calidad del modelo encontrado se llama *análisis de regresión*.

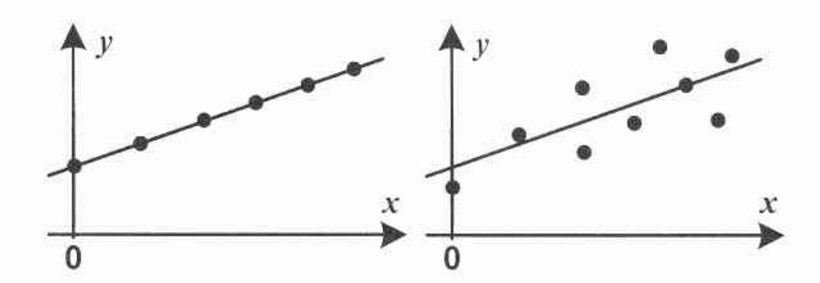


Figura 10.1: Modelos determinista y probabilístico.

Destaquemos algunas de las más importantes aplicaciones del análisis de regresión:

1. Descripción cuantitativa de las relaciones entre una variable dada y un conjunto de variables.
2. Interpolación entre valores de una función.
3. Predicción y pronóstico de datos.

En lo que sigue, nuestro interés será determinar una ecuación que relacione una variable dada con otra variable de respuesta, bajo el supuesto que ellas se vinculan mediante una ecuación lineal de primer grado, caso particular conocido como *regresión lineal simple*.

10.2. Modelo lineal simple

Recordemos que la ecuación de la recta es

$$y = \beta_0 + \beta_1 x,$$

donde β_0 es la intercepción de la recta con el eje y y β_1 es la pendiente de la recta. (Véase Figura 10.2)

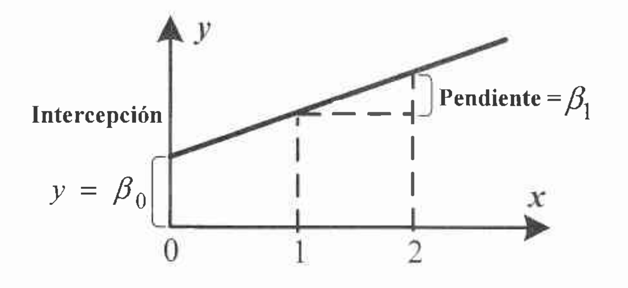


Figura 10.2: Recta de ecuación $y = \beta_0 + \beta_1 x$.

Este modelo es determinista porque no considera el error y los valores de y se obtienen, de manera exacta, al sustituir los valores de x en la ecuación de la recta.

Cuando se desea incorporar al modelo determinista el efecto aleatorio de las variables se le añade una componente que corresponde al error y el modelo queda como

$$y = \beta_0 + \beta_1 x + \varepsilon, \tag{10.1}$$

donde

y es la variable a ser modelada o *variable dependiente.*

x es la variable que se usa como predictor de y o *variable independiente.*

ε el componente aleatorio del error.

β_0 la ordenada del punto donde la recta interseca al eje y.

β_1 la pendiente de la recta.

A la variable independiente, también, se le denomina *predictora* y a la variable dependiente se llama *respuesta.*

Para recoger el efecto aleatorio del error, haremos las siguientes hipótesis sobre ε:

1. Se distribuye normalmente con media cero y varianza σ^2: $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.
2. Los errores, correspondientes a dos observaciones distintas, son independientes entre sí: $\mathbf{E}(\varepsilon_i \varepsilon_j) = 0$.

Ejemplos de modelos de regresión que se presentan en la vida cotidiana son los siguientes:

y	x	ε
Presupuesto de gastos de un hogar	Número de miembros del hogar	Efecto del nivel socioeconómico, tenencia de la vivienda, servicios que dispone, etc.
Precio de un departamento	Área de construcción	Efecto de la zona de ubicación, tipo de acabados, piso en el que se encuentra, etc.
Crecimiento anual de un árbol	Edad del árbol	Efecto de las variaciones climáticas, variedad del árbol, fertilización de la tierra, etc.
Precio de un libro	Número de páginas del libro	Efecto del tipo de papel, la encuadernación, número de ilustraciones, etc.

En el análisis de regresión es necesario tener en cuenta los siguientes pasos que llevan a estimar un buen modelo, que se ajuste a los datos:

1. Tener una visión clara de los objetivos del estudio, para determinar cuál ha de ser la variable respuesta y qué variables pueden incluirse como variables independientes.
2. Recopilar los datos correspondientes a las variables identificadas como dependiente e independientes.

3. Postular un modelo, al que se supone se ajustan los datos (en nuestro caso se presume que es el lineal simple).
4. Determinar la ecuación de regresión; es decir, estimar los coeficientes del modelo propuesto.
5. Comprobar estadísticamente la adecuación del modelo. Esto incluye la realización de pruebas estadísticas sobre los parámetros, la ejecución de transformaciones de las variables para obtener un mejor ajuste o retirar variables de una ecuación si su aporte no es significativo en la ecuación de predicción.
6. Cuando la ecuación sea satisfactoria, usar el modelo para efectuar estimaciones o predicciones.

Una vez que se han cumplido los tres primeros pasos, nuestro objetivo será estimar los coeficientes del modelo y comprobar la adecuación del modelo.

10.3. Método de los mínimos cuadrados

Para estimar los coeficientes de la ecuación de regresión se empleará el *método de los mínimos cuadrados*, consistente en minimizar la suma de los cuadrados de los errores; esto es, que si se nota a la ecuación de predicción por

$$\widehat{y} = b_0 + b_1 x, \tag{10.2}$$

donde b_0 y b_1 son los estimadores de β_0 y β_1, respectivamente; ellos deben ser tales que la suma de los cuadrados de las diferencias entre los valores observados de la variable respuesta y su estimación por la ecuación de regresión sea mínima.

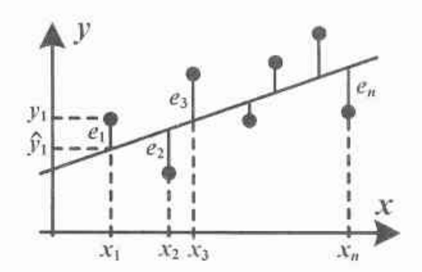


Figura 10.3: Interpretación del método de los mínimos cuadrados.

Si se dispone de n pares de observaciones de las variables independiente y dependiente $(x_1; y_1)$, $(x_2; y_2)$, ..., $(x_n; y_n)$ y si $\widehat{y}_i$ son los valores de las predicciones de y:

$$\widehat{y}_i = b_0 + b_1 x_i.$$

Entonces, los residuos de la predicción (errores) se calculan por $e_i = y_i - \widehat{y}_i$.

Nosotros buscaremos los valores de b_0 y b_1 que minimicen la suma de los cuadrados de los errores, también llamada *suma de los cuadrados de los residuos*:

$$\begin{aligned} SCE &= \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \\ &= \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2. \end{aligned}$$

Derivando SCE con respecto a b_0 y b_1, e igualando el resultado a cero se obtienen las ecuaciones:

$$\begin{aligned}\frac{\partial(SCE)}{\partial b_0} &= -2\sum_{i=1}^{n}(y_i - b_0 - b_1x_i) = 0,\\ \frac{\partial(SCE)}{\partial b_1} &= -2\sum_{i=1}^{n}x_i(y_i - b_0 - b_1x_i) = 0,\end{aligned}$$

cuya solución es

$$b_1 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = \frac{SC_{xy}}{SC_{xx}}, \qquad (10.3)$$

$$b_0 = \overline{y} - b_1\overline{x}, \qquad (10.4)$$

donde $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ y $\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ son los promedios de los valores de las variables independiente y dependiente.

Una vez obtenidos los valores de b_0 y b_1 se los sustituye en la ecuación; de esta manera queda establecida la recta de predicción por mínimos cuadrados:

$$\widehat{y} = b_0 + b_1x.$$

Observación. En la estimación de los parámetros se debe tener presente la incorporación de errores de redondeo en el cálculo de SC_{xy} y de SC_{xx}; se recomienda el empleo de un número suficiente de cifras significativas al realizar los cálculos de forma manual.

Ejemplo. En un estudio para determinar la relación entre el peso de los automóviles y su consumo de combustible se escogió una muestra de 10 carros, con los siguientes resultados:

Consumo (l/100 km)	8	16	6	7	7	9	11	12	18	20
Peso (kg)	739	1187	655	729	888	797	963	802	1551	1650

Determinar la ecuación de regresión lineal simple.

Solución: Primero, establezcamos que la variable independiente es el peso y la dependiente es el consumo.

Para simplificar los cálculos, resumamos los componentes en una tabla:

$$\overline{x} = 996.1; \qquad \overline{y} = 11.4.$$

x_i	y_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})^2$	$(x_i - \overline{x})(y_i - \overline{y})$
739	8	−257.1	−3.4	66 100.41	874.14
1187	16	190.9	4.6	36 442.81	878.14
655	6	−341.1	−5.4	116 349.21	1841.94
729	7	−267.1	−4.4	71 342.41	1175.24
888	7	−108.1	−4.4	11 685.61	475.64
797	9	−199.1	−2.4	39 640.81	477.84
963	11	−33.1	−0.4	1095.61	13.24
802	12	−194.1	0.6	37 674.81	−116.46
1551	18	554.9	6.6	307 914.01	3662.34
1650	20	653.9	8.6	427 585.21	5623.54
Suma				1 115 830.9	14 905.6

Sustituyendo los valores de las sumas en las fórmulas (10.3) y (10.4):

$$\begin{aligned} b_1 &= \frac{SC_{xy}}{SC_{xx}} = \frac{14\,905.6}{1\,115\,830.9} = 0.0134, \\ b_0 &= \overline{y} - b_1\overline{x} = 11.4 - 0.0134 \times 996.1 = -1.9477. \end{aligned}$$

Entonces, la recta de ajuste por el método de los mínimos cuadrados es

$$\widehat{y} = -1.9477 + 0.0134x.$$

Ahora se puede, por ejemplo, predecir el consumo de un auto que pesa 1000 kg; esto es, $x = 1000$:

$$\widehat{y} = -1.9477 + 0.0134 \times 1000 = 11.45.$$

10.4. Comprobación de la adecuación del modelo

Una vez que se halla una estimación de la recta de regresión, es necesario determinar si la ecuación obtenida es un buen modelo para los datos y cuantificar el error que se comete cuando se emplea tal ecuación. Ésto se logra mediante el empleo de los coeficientes de correlación y de determinación, y a través de la realización de pruebas estadísticas sobre los parámetros.

10.4.1. El coeficiente de correlación

Recordemos que si se tienen dos variables aleatorias, una medida de la relación que existe entre ellas es el coeficiente de correlación ρ. Análogamente, para determinar si existe una relación lineal entre las variables predictora y de respuesta se utiliza el *coeficiente de correlación lineal de Pearson*, denotado por r, que se define por

$$r = \frac{SC_{xy}}{\sqrt{SC_{xx}\,SC_{yy}}} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right)}}.$$

El coeficiente de correlación tiene las siguientes propiedades:

1. El rango de variación de r está entre -1 y 1, siendo su signo el mismo que el de b_1.
2. Un valor de r cercano a cero indica que no existe o hay poca relación (lineal) entre x y y.
3. Valores de r cercanos a 1 o a -1 indican que existe una fuerte relación entre las dos variables.
4. Si $r = 1$ o $r = -1$, todos los valores caen exactamente en la recta y se tiene un modelo determinístico.

En la Figura 10.4 se muestran diversos diagramas de la relación entre x y y, según los valores de r.

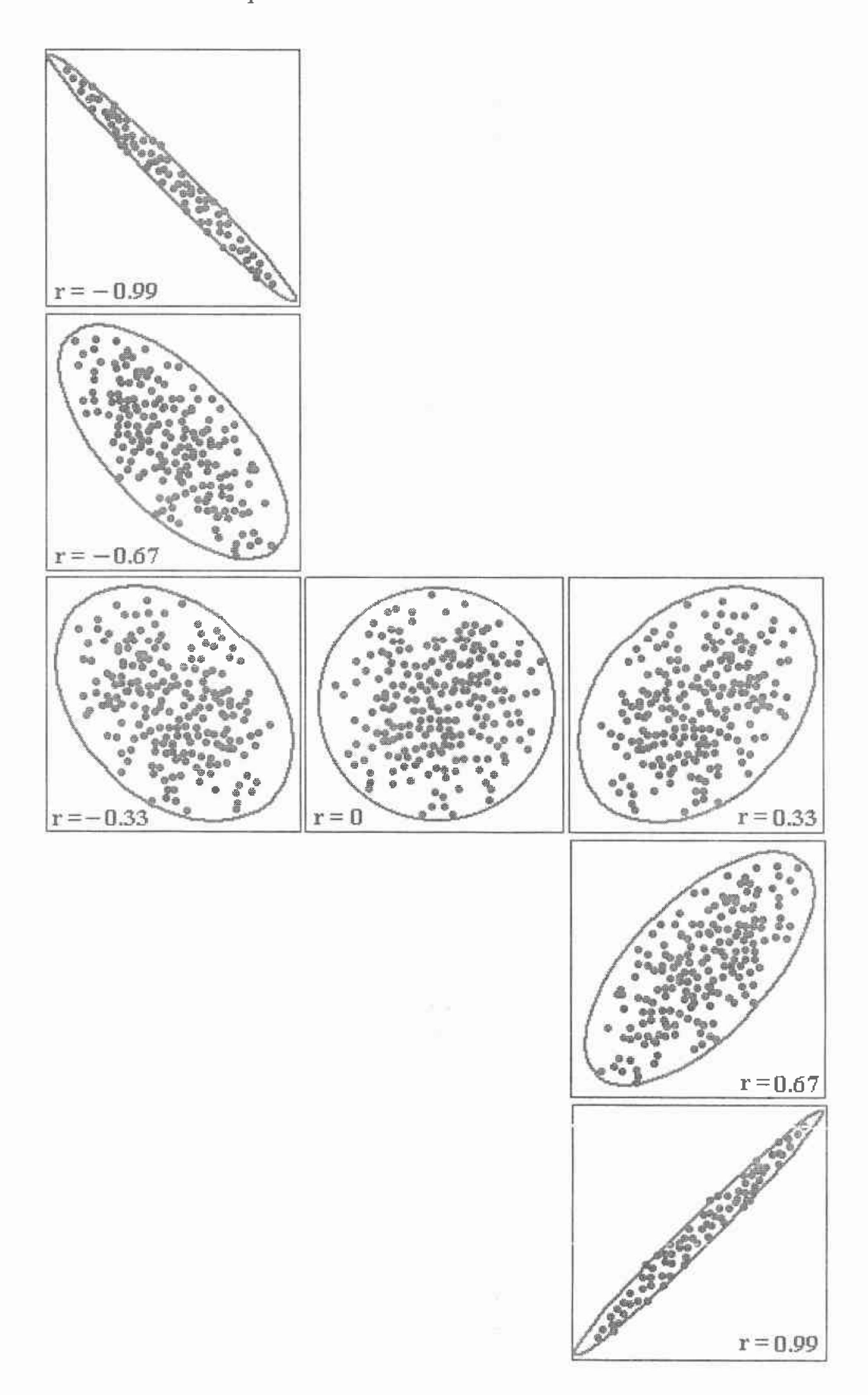


Figura 10.4: Diagramas de dispersión de los datos y valores aproximados del coeficiente de correlación.

Siempre debemos tener en cuenta que el coeficiente r solo aporta información cuando existe una relación lineal entre las variables. Pudiera suceder que se tenga un valor de r cercano a 0 y, sin embargo, haya una relación no-lineal entre las variables.

Como el coeficiente r es un estimador de ρ, se pueden efectuar pruebas sobre la significación del coeficiente de correlación muestral.

Prueba de hipótesis sobre ρ

1. *Hipótesis Nula.* H_0: $\rho = 0$.
2. *Hipótesis Alternativa.* H_1: $\rho \neq 0$.
3. *Estadístico de Prueba.* $t_{obs} = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$.
4. *Región de Rechazo.* Se rechaza H_0 si $t_{obs} < -t_{\alpha/2}(n-2)$ o $t_{obs} > t_{\alpha/2}(n-2)$.

También, se pueden conducir pruebas unilaterales sobre ρ, pero ellas solo tienen un valor estadístico y su valor práctico es restringido.

Ejemplo (Continuación). Para los datos del consumo de gasolina de los automóviles: a) calcular el coeficiente de correlación de Pearson; b) realizar una prueba para probar si $\rho = 0$ al nivel de significación de 0.05.

Solución: Los componentes de r son: $SC_{xx} = 1\,115\,830.9$, $SC_{yy} = 224.4$, $SC_{xy} = 14\,905.6$.

a) El coeficiente de correlación es

$$r = \frac{SC_{xy}}{\sqrt{SC_{xx}\,SC_{yy}}} = \frac{14\,905.6}{\sqrt{1\,115\,830.9 \times 224.4}} = 0.94197.$$

Como r es muy cercano a 1, existe una fuerte relación lineal entre las variables.

b) Para la prueba bilateral consideremos un nivel de significación del 5 %.

1. *Hipótesis Nula.* H_0: $\rho = 0$.
2. *Hipótesis Alternativa.* H_1: $\rho \neq 0$.
3. *Estadístico de Prueba.* $t_{obs} = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \dfrac{(0.94197)\sqrt{10-2}}{\sqrt{1-(0.94197)^2}} = 7.94.$
4. *Región de Rechazo.* Tenemos que $t_{0.025}(8) = 2.306$ y la región es $|t_{obs}| > 2.306$.
5. *Decisión.* Como $7.94 > 2.306$, se rechaza H_0; es decir, hay evidencia que indica que existe correlación entre el consumo de combustible y el peso del carro.

10.4.2. El coeficiente de determinación

Otra medida de la relación entre las variables es el *coeficiente de determinación*, r^2. Su empleo se debe a que da mayor fuerza de interpretación a la relación entre las variables. Tenemos que

$$SC_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2,$$

que se le conoce como *suma de los cuadrados alrededor de la media* de y o *suma de cuadrados corregida* de y.

Si denotamos por:

- $SCE = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$, que se denomina como *suma de cuadrados de los errores.* Este término también se puede calcular mediante la relación

$$SCE = SC_{yy} - b_1 SC_{xy} = SC_{yy} - \frac{(SC_{xy})^2}{SC_{xx}}.$$

- $SCR = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$, se designa como *suma de los cuadrados debido a la regresión.*

Se puede demostrar que

$$\begin{aligned} SC_{yy} &= SCE + SCR, \\ \sum_{i=1}^{n}(y_i - \overline{y})^2 &= \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2. \end{aligned}$$

Entonces, tenemos que

$$\begin{aligned} r^2 &= \frac{(SC_{xy})^2}{SC_{xx}\,SC_{yy}} = \frac{SC_{yy} - SCE}{SC_{yy}} \\ &= 1 - \frac{SCE}{SC_{yy}}. \end{aligned}$$

También, se puede expresar como

$$r^2 = \frac{SC_{yy} - SCE}{SC_{yy}} = \frac{SCR}{SC_{yy}}.$$

Interpretación. El coeficiente r^2 toma valores entre 0 y 1 y se puede interpretar de varias maneras:

1. **Como medida de reducción el error total.** Cuando $SCE = 0$, $r^2 = 1$ y cuando $SCR = 0$, $r^2 = 0$ Luego, r^2 representa la reducción relativa en la suma de SC_{yy} cuando se ha ajustado una recta de regresión.

   Por ejemplo, un valor de $r^2 = 0.7$ significa que el 70 % de la variación total es explicada por la recta que ajusta los datos y el 30 % restante es atribuible al error de ajuste.

2. **Como medida de bondad de ajuste.** Cuando el ajuste es perfecto, $r^2 = 1$ y cuando no hay ajuste, $r^2 = 0$.

   Cuanto mayor sea el valor de r^2, mejor será el ajuste y mayor utilidad tendrá como instrumento de predicción.

3. **Como medida de la linealidad de los puntos.** Cuando r^2 se aproxima a 1, el gráfico de los datos se acerca a una línea recta. Si la relación entre dos variables no es lineal, $r^2 = 0$.

Ejemplo (Continuación). Para los datos del consumo de gasolina de los automóviles, calcular el coeficiente de determinación.

Solución: Se tiene que

$$\begin{aligned} SCE &= SC_{yy} - \frac{(SC_{xy})^2}{SC_{xx}} = 224.4 - \frac{(14\,905.6)^2}{1\,115\,830.9} \\ &= 25.287. \end{aligned}$$

De modo que

$$r^2 = 1 - \frac{SCE}{SC_{yy}} = 1 - \frac{25.287}{224.4} = 0.8873.$$

Es decir, el 88.73 % de la variabilidad de los datos es recogida por la recta de regresión.

10.4.3. Estimación de la varianza de la regresión

Al proponer el modelo probabilístico

$$y = \beta_0 + \beta_1 x + \varepsilon$$

se estableció que el error aleatorio está distribuido según una ley normal de media cero y varianza σ^2. Entonces, cada valor observado de y está influido por tal error. Además, a los errores asociados con distintas observaciones se los considerará mutuamente independientes.

La varianza σ^2 del error es desconocida y su estimador insesgado es

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{n-2} = \frac{SCE}{n-2}.$$

A s, que es la desviación estándar de la regresión muestral, también se la denomina *error estándar de estimación*.

Ejemplo (Continuación). Calcular la estimación de σ^2 para los datos del consumo de combustible de los carros.

Solución: Anteriormente se obtuvo que $SCE = 25.287$. De modo que

$$s^2 = \frac{SCE}{n-2} = \frac{25.287}{10-2} = 3.161.$$

Aunque s^2 se puede considerar una medida de la calidad de ajuste, su principal utilidad se encuentra en la determinación de la bondad de ajuste, ya sea mediante un intervalo de confianza o con una prueba de hipótesis.

10.4.4. Inferencias acerca de la pendiente de la recta, β_1

En primer lugar se desea estudiar si existe o no existe relación entre las variables x y y. Se desea contestar a la pregunta ¿aporta x información para predecir y? Esta pregunta se refiere a β_1, pues afirmar que y no se relaciona linealmente con x equivale a decir que $\beta_1 = 0$.

Entonces, se desea probar la hipótesis nula «x no contribuye con información para predecir y»; contra la hipótesis alternativa, «las variables están relacionadas de forma lineal con una pendiente distinta de cero»; es decir,

$$H_0:\ \beta_1 = 0,$$
$$H_1:\ \beta_1 \neq 0.$$

Para efectuar la prueba habrá que encontrar la distribución de muestreo de b_1.

Distribución de muestreo de b_1

Si los componentes del error son variables aleatorias independientes normalmente distribuidas con media cero y varianza σ^2, la distribución de muestreo del estimador b_1 es normal con media β_1 y desviación estándar

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{SC_{xx}}}.$$

Ésto quiere decir que b_1 es un estimador insesgado para β_1, pues $\mathbf{E}(b_1) = \beta_1$ y que la desviación estándar de b_1 puede estimarse por

$$s_{b_1} = \frac{s}{\sqrt{SC_{xx}}},$$

donde s es el error estándar de la estimación.

Entonces, la variable aleatoria $t = \dfrac{b_1 - \beta_1}{s_{b_1}}$ sigue una ley t a $(n-2)$ grados de libertad.

Prueba de hipótesis acerca de β_1

a) Prueba unilateral.

1. *Hipótesis Nula.* H_0: $\beta_1 = 0$.
2. *Hipótesis Alternativa.* H_1: $\beta_1 < 0$ (o bien H_1: $\beta_1 > 0$).
3. *Estadístico de Prueba.* $t_{obs} = \dfrac{b_1}{s/\sqrt{SC_{xx}}}$.
4. *Región de Rechazo.* Se rechaza H_0 si $t_{obs} < -t_\alpha(n-2)$ (o bien $t_{obs} > t_\alpha(n-2)$ cuando H_1: $\beta_1 > 0$.)

b) Prueba bilateral.

1. *Hipótesis Nula.* H_0: $\beta_1 = 0$.
2. *Hipótesis Alternativa.* H_1: $\beta_1 \neq 0$.
3. *Estadístico de Prueba.* $t_{obs} = \dfrac{b_1}{s/\sqrt{SC_{xx}}}$.
4. *Región de Rechazo.* Se rechaza H_0 si $t_{obs} < -t_{\alpha/2}(n-2)$ o $t_{obs} > t_{\alpha/2}(n-2)$.

Intervalo de confianza de $100(1-\alpha)\,\%$ para β_1

Un intervalo de confianza para la pendiente β_1, a nivel $100(1-\alpha)\,\%$ es

$$\left(b_1 - t_{\alpha/2}(n-2)\frac{s}{\sqrt{SC_{xx}}};\ b_1 + t_{\alpha/2}(n-2)\frac{s}{\sqrt{SC_{xx}}}\right).$$

Ejemplo (Continuación). Para los datos del consumo de combustible de varios carros: a) probar si $\beta_1 = 0$, a un nivel de significación de 0.05; b) obtener el intervalo de confianza al 95 %.

Solución: Antes se había determinado los valores de s y de SC_{xx}:

$$s = \sqrt{3.161} = 1.7779, \qquad SC_{xx} = 1\,115\,830.9.$$

a) Realizaremos una prueba bilateral:

1. *Hipótesis Nula.* H_0: $\beta_1 = 0$.
2. *Hipótesis Alternativa.* H_1: $\beta_1 \neq 0$.
3. *Estadístico de Prueba.* $t_{obs} = \dfrac{b_1}{s/\sqrt{SC_{xx}}} = \dfrac{0.0134}{1.7779/\sqrt{1\,115\,830.9}} = 7.9615.$
4. *Región de Rechazo.* Como $t_{obs} > t_{0.025}(8) = 2.306$ y la región es $|t_{obs}| > 2.306$
5. *Decisión.* Se rechaza la hipótesis nula; es decir, el consumo de combustible aumenta a medida que se incrementa el peso de los carros.

b) El intervalo de confianza es:

$$\left(b_1 - t_{\alpha/2}(n-2)\frac{s}{\sqrt{SC_{xx}}};\, b_1 + t_{\alpha/2}(n-2)\frac{s}{\sqrt{SC_{xx}}}\right) =$$
$$\left(0.0134 - 2.306\frac{1.7779}{\sqrt{1\,115\,830.9}};\, 0.0134 + 2.306\frac{1.7779}{\sqrt{1\,115\,830.9}}\right) = (0.0095; 0.0173).$$

10.4.5. Inferencias acerca de la intercepción de la recta, β_0

Supongamos que se desea averiguar si β_0 es igual, a un valor específico; ello se logra efectuando una prueba de hipótesis o con un intervalo de confianza, de manera similar a la descrita para β_1.

Se puede demostrar que la desviación estándar de b_0 es

$$\sigma_{b_0} = \sigma\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n\,SC_{xx}}}.$$

El estadístico $t = \dfrac{(b_0 - \beta_{00})\sqrt{n\,SC_{xx}}}{s\sqrt{\sum_{i=1}^{n} x_i^2}}$ sigue una distribución t con $(n-2)$ grados de libertad.

Prueba de hipótesis acerca de β_0

1. *Hipótesis Nula.* H_0: $\beta_0 = \beta_{00}$.
2. *Hipótesis Alternativa.* H_1: $\beta_0 \neq \beta_{00}$.
3. *Estadístico de Prueba.* $t_{obs} = \dfrac{b_0 - \beta_{00}}{s}\sqrt{\dfrac{n\,SC_{xx}}{\sum_{i=1}^{n} x_i^2}}.$
4. *Región de Rechazo.* Se rechaza H_0 si $t_{obs} < -t_{\alpha/2}(n-2)$ o $t_{obs} > t_{\alpha/2}(n-2)$.

Intervalo de confianza de $100(1-\alpha)\%$ para β_0

$$\left(b_0 - t_{\alpha/2}(n-2)\frac{s\sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{n\,SC_{xx}}};\, b_0 + t_{\alpha/2}(n-2)\frac{s\sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{n\,SC_{xx}}}\right).$$

Ejemplo (Continuación). Para los datos del consumo de combustible de varios carros: a) probar si $\beta_0 = 0$, a un nivel de significación de 0.05; b) obtener el intervalo de confianza al 95 %.

Solución:

a) Realizamos una prueba bilateral.

1. *Hipótesis Nula.* H_0: $\beta_0 = 0$.
2. *Hipótesis Alternativa.* H_1: $\beta_0 \neq 0$.
3. *Estadístico de Prueba.*

$$t_{obs} = \frac{b_0 - \beta_{00}}{s}\sqrt{\frac{n\,SC_{xx}}{\sum\limits_{i=1}^{n} x_i^2}} = \frac{-1.9477 - 0}{1.7779}\sqrt{\frac{10 \times 1\,115\,830.9}{11\,037\,983}} = -1.101.$$

4. *Región de Rechazo.* La región crítica es $|t_{obs}| > t_{0.025}(8) = 2.306$.
5. *Decisión.* Como $|-1.101| < 2.306$, no se rechaza H_0; es decir, β_0 pudiera ser igual a cero.

b) El intervalo de confianza es

$$\left(b_0 - t_{\alpha/2}(n-2)\frac{s\sqrt{\sum\limits_{i=1}^{n} x_i^2}}{\sqrt{n\,SC_{xx}}}; b_0 + t_{\alpha/2}(n-2)\frac{s\sqrt{\sum\limits_{i=1}^{n} x_i^2}}{\sqrt{n\,SC_{xx}}} \right) =$$

$$\left(-1.9477 - 2.306\frac{1.7779\sqrt{11\,037\,983}}{\sqrt{10 \times 1\,115\,830.9}}; -1.9477 + 2.306\frac{1.7779\sqrt{11\,037\,983}}{\sqrt{10 \times 1\,115\,830.9}} \right) = (-6.025; 2.130).$$

10.5. Análisis de la varianza

El procedimiento del *análisis de la varianza* (ANOVA) es una metodología estadística que permite la comparación de dos o más medias poblacionales midiendo la variación dentro de las muestras. En nuestro caso emplearemos el análisis de la varianza para efectuar pruebas sobre los parámetros estimados del modelo y así conocer su exactitud.

Anteriormente establecimos que se cumple la siguiente igualdad:

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \overline{y})^2 &= \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 \\ SC_{yy} &= SCE + SCR. \end{aligned} \tag{10.5}$$

Significa que la suma de los cuadrados corregida es igual a la suma de cuadrados de los errores más la suma de cuadrados debida a la regresión.

La suma de los cuadrados corregida tiene $(n-1)$ grados de libertad, la suma de cuadrados de los errores tiene $(n-2)$ grados de libertad y la suma de cuadrados debida a la regresión tiene 1 grado de libertad. Es decir, la igualdad correspondiente a los grados de libertad de la ecuación (10.5) es

$$n - 1 = (n-2) + 1 \tag{10.6}$$

De las ecuaciones (10.5) y (10.6) se tiene la tabla de análisis de la varianza, como la que se presenta a continuación.

Tabla de Análisis de la Varianza				
Fuente de variación	**Grados de libertad** ($g.l.$)	**Suma de cuadrados** (SC)	**Cuadrado medio** (MC)	F
Regresión	1	SCR	$MCR = SCR/1$	$F_{obs} = \frac{MCR}{s^2}$
Error o residual	$n-2$	SCE	$s^2 = SCE/(n-2)$	
Total corregido	$n-1$	SC_{yy}		

La columna del «cuadrado medio» (MC) se obtiene al dividir cada una de las suma de los cuadrados entre sus correspondientes grados de libertad.

El valor de F_{obs} resulta de la división del cuadrado medio de la regresión para el cuadrado medio residual: $F_{obs} = \frac{MCR}{s^2}$.

Una vez elaborada la tabla de análisis de varianza, el valor de F_{obs} se emplea para conducir una prueba de hipótesis sobre la razón de dos varianzas, que sirve para probar si $\beta_1 = 0$. La prueba es la siguiente:

1. *Hipótesis Nula.* H_0: $\beta_1 = 0$.
2. *Hipótesis Alternativa.* H_1: $\beta_1 \neq 0$.
3. *Estadístico de Prueba.* $F_{obs} = \frac{MCR}{s^2}$.
4. *Región de Rechazo.* Se rechaza H_0 si $F_{obs} > F_\alpha(1, n-2)$.

Ejemplo (Continuación). Para los datos del consumo de combustible de varios carros: a) realizar la tabla de análisis de la varianza; b) probar si $\beta_1 = 0$, a un nivel de significación de 0.05.

Solución: Las sumas de los cuadrados correspondientes son

$$SC_{yy} = 224.4, \quad SCE = 25.287, \quad SCR = 199.113.$$

a) La tabla de análisis de la varianza queda como sigue:

Tabla de Análisis de la Varianza				
Fuente de variación	Grados de libertad (g.l.)	Suma de cuadrados (SC)	Cuadrado medio (MC)	F
Regresión	1	199.113	199.113	62.99
Error	8	25.287	3.1609	
Total corregido	9	224.4		

b) Realizamos la prueba de hipótesis:

1. *Hipótesis Nula.* H_0: $\beta_1 = 0$.
2. *Hipótesis Alternativa.* H_1: $\beta_1 \neq 0$.
3. *Estadístico de Prueba.* $F_{obs} = 62.99$.
4. *Región de Rechazo.* Como $F_{obs} > F_{0.05}(1, 8) = 5.32$, la región es $F_{obs} > 5.32$.
5. *Decisión.* Se rechaza la hipótesis nula. Es más, por la magnitud del estadístico de prueba, se deduce que el nivel de significación del contraste es sumamente bajo.

10.6. Uso del modelo para estimación y predicción

Una vez que el modelo es satisfactorio habrá que emplearlo para realizar estimaciones y predicciones que servirán para analizar el comportamiento de la variable respuesta ante condiciones que no fueron probadas empíricamente.

Previamente establezcamos las diferencia entre intervalos de estimación y de predicción:

1. Un intervalo de estimación se utiliza para estimar la media de Y, dado un valor de x; es decir, estimar el valor $\mu_{Y|x_0}$ mediante un intervalo, cuando $x = x_0$. Un investigador puede desear conocer el consumo medio, para un peso dado del auto.

2. Un intervalo de predicción se utiliza para estimar un valor particular de Y, cuando $x = x_0$; es decir, estimar el valor Y_0 mediante un intervalo de confianza. El investigador puede desear conocer el consumo de un auto que tiene un peso dado.

Nótese que los valores de estimación y de predicción de Y son idénticos en los dos casos, la diferencia radica en la precisión relativa de cada una, que se ven reflejadas en sus varianzas e intervalos de confianza.

10.6.1. Intervalo de confianza para la estimación, $\mathbf{E(Y)}$, cuando $\mathbf{x = x}_p$

En el caso de la estimación se trata de calcular

$$\mathbf{E}(Y_p) = \beta_0 + \beta_1 x_p,$$

donde x_p es el valor particular de x para el cual se está haciendo la estimación. El estimador es

$$\widehat{y}_p = b_0 + b_1 x_p,$$

por tanto, las fuentes de error para estimar $\mathbf{E}(Y)$ son b_0 y b_1.

La varianza es

$$\sigma^2_{\widehat{y}} = \sigma^2 \left(\frac{1}{n} + \frac{(x_p - \overline{x})^2}{SC_{xx}} \right).$$

La varianza estimada correspondiente a $\widehat{y}$ emplea s^2 en lugar de σ^2 en la expresión anterior.

Un intervalo de confianza al $100(1-\alpha)\,\%$ para la estimación, $\mathbf{E}(Y_p)$, en el punto $\mathbf{x}_p$ es

$$\left(\widehat{y}_p - t_{\alpha/2}(n-2)s\sqrt{\frac{1}{n} + \frac{(x_p - \overline{x})^2}{SC_{xx}}}; \widehat{y}_p + t_{\alpha/2}(n-2)s\sqrt{\frac{1}{n} + \frac{(x_p - \overline{x})^2}{SC_{xx}}} \right).$$

10.6.2. Intervalo de confianza para la predicción, $\mathbf{y}_p$, cuando $\mathbf{x = x}_p$

En el caso de la predicción el objetivo es calcular

$$y_p = \beta_0 + \beta_1 x_p + \varepsilon_p,$$

el valor pronosticado cuando x_p es el valor particular de x. El estimador es

$$\widehat{y}_p = b_0 + b_1 x_p,$$

pero las fuentes del error de predicción son b_0, b_1 y ε_p.

La varianza es

$$\sigma^2_{(y-\widehat{y})} = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{SC_{xx}}\right)$$

y en su estimación se emplea s^2 en lugar de σ^2 en la expresión anterior.

El intervalo de confianza de nivel $100(1-\alpha)\,\%$ para la predicción, y_p, es:

$$\left(\widehat{y}_p - t_{\alpha/2}(n-2)s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{SC_{xx}}}; \widehat{y}_p + t_{\alpha/2}(n-2)s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{SC_{xx}}}\right).$$

Ejemplo (Continuación). Para los datos del consumo de combustible de los carros, obtener los intervalos de estimación y de predicción para un carro cuyo peso es 1000 kg.

Solución: Como se indicó los valores estimado y pronosticado son los mismos:

$$\widehat{y}_p = b_0 + b_1 x_p = -1.9477 + 0.0134 \times 1000 = 11.45.$$

La varianza $\sigma^2_{\widehat{y}}$ es

$$\begin{aligned}\sigma^2_{\widehat{y}} &= \sigma^2\left(\frac{1}{n} + \frac{(x_p - \overline{x})^2}{SC_{xx}}\right)\\ &= 3.161\left(\frac{1}{10} + \frac{(1000 - 996.1)^2}{1\,115\,830.9}\right) = 0.316.\end{aligned}$$

y la varianza $\sigma^2_{(y-\widehat{y})}$ es

$$\begin{aligned}\sigma^2_{(y-\widehat{y})} &= \sigma^2\left(1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{SC_{xx}}\right)\\ &= 3.161\left(1 + \frac{1}{10} + \frac{(1000 - 996.1)^2}{1\,115\,830.9}\right) = 3.477.\end{aligned}$$

El intervalo de estimación correspondiente es

$$\begin{aligned}I_e &= \left(11.45 - 2.306\sqrt{0.316}; 11.45 + 2.306\sqrt{0.316}\right)\\ &= (10.15; 12.75).\end{aligned}$$

El intervalo de predicción es

$$\begin{aligned}I_p &= \left(11.45 - 2.306\sqrt{3.477}; 11.45 + 2.306\sqrt{3.477}\right)\\ &= (7.15; 15.75).\end{aligned}$$

Como se ve, el intervalo de predicción es más ancho que el de estimación. ◄

Se deberá tener la precaución de no usar el modelo obtenido para estimar el valor medio de y o predecir un valor particular de y, con valores de x que se encuentren fuera del rango de la variable independiente que sirvió para su elaboración.

El modelo podría ajustarse perfectamente a los datos recogidos para hallar la ecuación, pero nada garantiza que el mismo comportamiento se consiga fuera de tales límites, pudiendo darse un ajuste bastante malo. (Figura 10.5.)

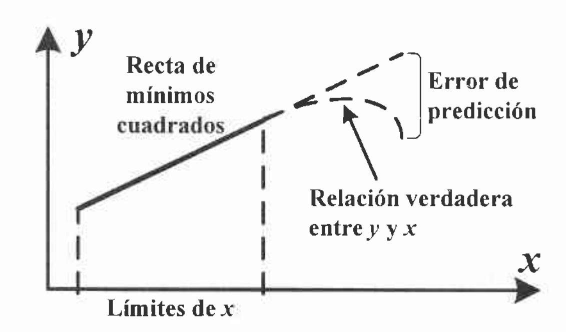


Figura 10.5: Uso errado de un modelo para realizar predicciones fuera del rango de definición de x.

10.7. Formulación matricial de la regresión lineal

Si se tiene n observaciones de las variables independiente y dependiente: $(x_1; y_1)$, $(x_2; y_2)$, ..., $(x_n; y_n)$ y la ecuación lineal que las relaciona es

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

Entonces, se tienen las siguientes igualdades:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \varepsilon_1, \\ y_2 &= \beta_0 + \beta_1 x_2 + \varepsilon_2, \\ \vdots &= \vdots \\ y_n &= \beta_0 + \beta_1 x_n + \varepsilon_n. \end{aligned} \tag{10.7}$$

Si se ponen las matrices

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

donde

Y es un vector $n \times 1$ de las observaciones.

X es una matriz $n \times 2$ de los valores de la variable independiente, cuya primera columna tiene la particularidad que todos sus componentes son iguales a 1.

β es un vector 2×1 de los parámetros de la ecuación.

ε es un vector $n \times 1$ de los errores aleatorios.

El sistema de ecuaciones (10.7) es equivalente a la igualdad

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Y si se nota como $\mathbf{b}$ al vector de los estimadores de los parámetros de regresión

$$\mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix},$$

se tiene el sistema de estimación de los parámetros $\widehat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$.

El vector de errores es igual a $\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbf{X}\mathbf{b}$.

Con ello se tiene[1]

$$\begin{aligned} \boldsymbol{\varepsilon}^t\boldsymbol{\varepsilon} &= (\mathbf{Y} - \mathbf{X}\mathbf{b})^t(\mathbf{Y} - \mathbf{X}\mathbf{b}) \\ &= \mathbf{Y}^t\mathbf{Y} - 2\mathbf{b}\mathbf{X}^t\mathbf{Y} + \mathbf{b}^t\mathbf{X}^t\mathbf{X}\mathbf{b}. \end{aligned}$$

Derivando esta última igualdad respecto a $\mathbf{b}$, e igualando el resultado a cero

$$\frac{\partial \boldsymbol{\varepsilon}^t\boldsymbol{\varepsilon}}{\partial \mathbf{b}} = -2\mathbf{X}^t\mathbf{Y} + 2\mathbf{X}^t\mathbf{X}\mathbf{b} = \mathbf{0},$$

cuyo resultado es el sistema de ecuaciones

$$\mathbf{X}^t\mathbf{X}\mathbf{b} = \mathbf{X}^t\mathbf{Y}. \qquad (10.8)$$

Las matrices $\mathbf{X}^t\mathbf{X}$ y $\mathbf{X}^t\mathbf{Y}$ son:

$$\mathbf{X}^t\mathbf{X} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} \quad \text{y} \quad \mathbf{X}^t\mathbf{Y} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}.$$

Si la matriz $\mathbf{X}^t\mathbf{X}$ es inversible, se llega a la ecuación de estimación de los parámetros

$$\mathbf{b} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}.$$

La matriz $(\mathbf{X}^t\mathbf{X})^{-1}$ es

$$(\mathbf{X}^t\mathbf{X})^{-1} = \frac{1}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}.$$

Con ésto, los valores ajustados $\widehat{\mathbf{Y}}$ se obtienen evaluando

$$\widehat{\mathbf{Y}} = \mathbf{X}\mathbf{b}.$$

[1]En lo que sigue se empleará la notación $\mathbf{A}^t$ para indicar la matriz o vector transpuesto de $\mathbf{A}$.

y la ecuación de predicción de un valor y_p correspondiente a un valor x_p dado es

$$y_p = \mathbf{x}_p^t \mathbf{b},$$

con $\mathbf{x}_p^t = (1; x_p)$.

Ejemplo (Continuación). Con los datos del consumo de combustible de varios carros, plantear la formulación matricial de la regresión lineal y encontrar la ecuación de regresión.

Solución: Se forman las siguientes matrices:

$$\mathbf{Y} = \begin{pmatrix} 8 \\ 16 \\ \vdots \\ 20 \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & 739 \\ 1 & 1187 \\ \vdots & \vdots \\ 1 & 1650 \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_9 \end{pmatrix}.$$

El modelo es $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

Calculemos $(\mathbf{X}^t\mathbf{X})^{-1}$ y $\mathbf{X}^t\mathbf{Y}$:

$$\begin{aligned}
\mathbf{X}^t\mathbf{X} &= \begin{pmatrix} 1 & 1 & \dots & 1 \\ 739 & 1187 & \dots & 1650 \end{pmatrix} \begin{pmatrix} 1 & 739 \\ 1 & 1187 \\ \vdots & \vdots \\ 1 & 1650 \end{pmatrix} = \begin{pmatrix} 10 & 9961 \\ 9961 & 11\,037\,983 \end{pmatrix}, \\
(\mathbf{X}^t\mathbf{X})^{-1} &= \frac{1}{11\,158\,309} \begin{pmatrix} 11\,037\,983 & -9961 \\ -9961 & 10 \end{pmatrix}, \\
\mathbf{X}^t\mathbf{Y} &= \begin{pmatrix} 1 & 1 & \dots & 1 \\ 739 & 1187 & \dots & 1650 \end{pmatrix} \begin{pmatrix} 8 \\ 16 \\ \vdots \\ 20 \end{pmatrix} = \begin{pmatrix} 114 \\ 128\,461 \end{pmatrix}.
\end{aligned}$$

Así,

$$\begin{aligned}
\mathbf{b} &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y} \\
&= \frac{1}{11\,158\,309} \begin{pmatrix} 11\,037\,983 & -9961 \\ -9961 & 10 \end{pmatrix} \begin{pmatrix} 114 \\ 128\,461 \end{pmatrix} = \begin{pmatrix} -1.906 \\ 0.0134 \end{pmatrix}.
\end{aligned}$$

Por tanto, la ecuación de regresión[2] es $\widehat{y} = -1.906 + 0.0134x$.

10.7.1. Matriz de covarianza entre b_0 y b_1

Recordemos que

$$\operatorname{Var}(b_1) = \frac{\sigma^2}{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2}, \qquad \operatorname{Var}(b_0) = \frac{\sigma^2 \sum\limits_{i=1}^{n} x_i^2}{n \sum\limits_{i=1}^{n}(x_i - \overline{x})^2}.$$

[2] La diferencia en el valor del término independiente entre las formulaciones algebraica y matricial, se debe al número de decimales utilizados en el cálculo de la formulación algebraica.

Además, la covarianza entre b_0 y b_1 es

$$\text{Cov}(b_0, b_1) = \text{Cov}[(\overline{y} - b_1\overline{x}), b_1] = \frac{-\overline{x}\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Entonces, la matriz de varianza-covarianza del vector $\mathbf{b}$ es

$$\text{Var}(\mathbf{b}) = \begin{pmatrix} \text{Var}(b_0) & \text{Cov}(b_0, b_1) \\ \text{Cov}(b_0, b_1) & \text{Var}(b_1) \end{pmatrix} = \begin{pmatrix} \dfrac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \overline{x})^2} & \dfrac{-\overline{x}\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \\ \dfrac{-\overline{x}\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} & \dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \end{pmatrix}.$$

Si sacamos fuera de la matriz el factor σ^2, resulta que

$$\text{Var}(\mathbf{b}) = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}.$$

La varianza σ^2 se estima mediante

$$s^2 = \frac{\mathbf{Y}^t\mathbf{Y} - \mathbf{b}^t\mathbf{X}^t\mathbf{Y}}{n-2}.$$

10.7.2. Formulación matricial del análisis de la varianza

Con el empleo de la formulación matricial de los componentes de la regresión se puede poner las sumas de los cuadrados como

$$\begin{aligned} SCE &= \mathbf{Y}^t\mathbf{Y} - \mathbf{b}^t\mathbf{X}^t\mathbf{Y}, \\ SCR &= \mathbf{b}^t\mathbf{X}^t\mathbf{Y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}, \\ SC_{yy} &= \mathbf{Y}^t\mathbf{Y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}. \end{aligned}$$

Así, el coeficiente de determinación, en formulación matricial, es igual a

$$r^2 = \frac{SCR}{SC_{yy}} = \frac{\mathbf{b}^t\mathbf{X}^t\mathbf{Y} - \dfrac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}{\mathbf{Y}^t\mathbf{Y} - \dfrac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}.$$

También, se puede reconstruir la tabla de análisis de la varianza en su forma matricial:

Tabla de Análisis de la Varianza				
Fuente de variación	Grados de libertad (g.l.)	Suma de cuadrados (SC)	Cuadrado medio (MC)	F
Regresión	1	$SCR = \mathbf{b}^t\mathbf{X}^t\mathbf{Y} - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$	MCR	$F_{obs} = \frac{MCR}{s^2}$
Error o residual	$n-2$	$SCE = \mathbf{Y}^t\mathbf{Y} - \mathbf{b}^t\mathbf{X}^t\mathbf{Y}$	$s^2 = \frac{SCE}{n-2}$	
Total corregido	$n-1$	$SC_{yy} = \mathbf{Y}^t\mathbf{Y} - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$		

Ejemplo (Continuación). Con los datos del consumo de combustible, construir la tabla de análisis de la varianza.

Solución: Calculemos cada uno de los términos que intervienen en la tabla.

$$\begin{aligned}
\mathbf{b}^t\mathbf{X}^t\mathbf{Y} &= \begin{pmatrix} -1.906 & 0.0134 \end{pmatrix} \begin{pmatrix} 114 \\ 128\,461 \end{pmatrix} = 1498.71, \\
\mathbf{Y}^t\mathbf{Y} &= \begin{pmatrix} 8 & 16 & \cdots & 20 \end{pmatrix} \begin{pmatrix} 8 \\ 16 \\ \vdots \\ 20 \end{pmatrix} = 1524, \\
\sum_{i=1}^{n} y_i &= 114.
\end{aligned}$$

Entonces,

$$\begin{aligned}
SCR &= \mathbf{b}^t\mathbf{X}^t\mathbf{Y} - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 = 1498.71 - \frac{1}{10}(114)^2 = 199.11, \\
SCE &= \mathbf{Y}^t\mathbf{Y} - \mathbf{b}^t\mathbf{X}^t\mathbf{Y} = 1524 - 1498.71 = 25.29, \\
SC_{yy} &= \mathbf{Y}^t\mathbf{Y} - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 = 1524 - \frac{1}{10}(114)^2 = 224.4.
\end{aligned}$$

Con todo lo anterior, se tiene la tabla:

Tabla de análisis de la varianza				
Fuente de variación	Grados de libertad (g.l.)	Suma de cuadrados (SC)	Cuadrado medio (MC)	F
Regresión	1	199.11	199.11	62.99
Error	8	25.29	3.16	
Total corregido	9	224.40		

Obviamente, la tabla coincide con la elaborada mediante la formulación algebraica.

10.8. Transformación de modelos no-lineales a lineales

En muchas ocasiones los modelos no lineales pueden ser tratados como lineales si se efectúan algunas transformaciones a las variables, ya sea a la predictora, a la respuesta o a ambas.

Al emplear tales transformaciones se deberá tener la precaución de verificar que el modelo modificado cumple con la hipótesis sobre la distribución que siguen los errores.

1. **Modelo exponencial** (Figura 10.6.)

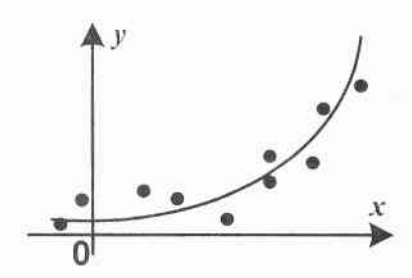


Figura 10.6: Modelo exponencial $y = e^{\beta_0+\beta_1 x+\varepsilon}$.

El modelo es

$$y = e^{\beta_0+\beta_1 x+\varepsilon}.$$

Tomando logaritmos en ambos miembros,

$$\ln y = \beta_0 + \beta_1 x + \varepsilon.$$

Si ponemos $z = \ln y$ queda $z = \beta_0 + \beta_1 x + \varepsilon$, que es un modelo lineal simple, que se estima con $\widehat{z} = b_0 + b_1 x$.

2. **Modelo recíproco o inverso** (Figura 10.7.)

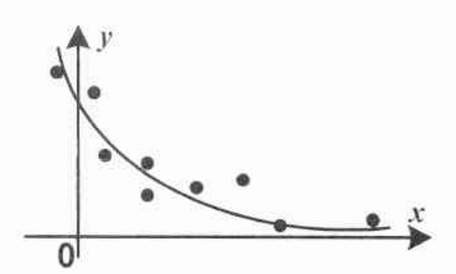


Figura 10.7: Modelo recíproco $y = \dfrac{1}{\beta_0 + \beta_1 x + \varepsilon}$.

El modelo recíproco o inverso es

$$y = \frac{1}{\beta_0 + \beta_1 x + \varepsilon}.$$

Tomando los inversos en ambos miembros,

$$\frac{1}{y} = \beta_0 + \beta_1 x + \varepsilon.$$

Poniendo $z = \dfrac{1}{y}$ la última igualdad queda como

$$z = \beta_0 + \beta_1 x + \varepsilon,$$

que es un modelo lineal usual.

3. **Modelo multiplicativo o potencial** (Figura 10.8.)

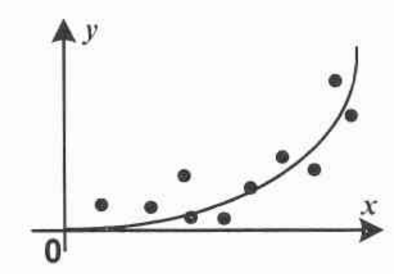


Figura 10.8: Modelo multiplicativo $y = \alpha x^{\lambda}\varepsilon$.

El modelo multiplicativo, también conocido como potencial, es

$$y = \alpha x^{\lambda}\varepsilon.$$

Tomando logaritmos en ambos miembros de la igualdad:

$$\ln y = \ln \alpha + \lambda \ln x + \ln \varepsilon.$$

Haciendo $z = \ln y$, $t = \ln x$, $\beta_0 = \ln \alpha$ y $\beta_1 = \lambda$, el modelo se escribe como

$$z = \beta_0 + \beta_1 t + \varepsilon,$$

que se estima por el modelo lineal $\widehat{z} = b_0 + b_1 t$.

Otros modelos no lineales comunmente utilizados son los siguientes:

4. **Logarítmico.** $y = \alpha + \beta \ln x$, en el que se utiliza la transformación $t = \ln x$.
5. **Compuesto.** $y = \alpha\,\beta^x$, que se linealiza mediante $\ln y = \ln \alpha + \ln \beta\, x$.
6. **Curva S.** $y = e^{\alpha + \beta/x}$, que se transforma en $\ln y = \alpha + \beta t$, con $t = \dfrac{1}{x}$.

Se recomienda que el lector realice las operaciones necesarias para linealizar estos modelos.

Ejemplo. En el desarrollo de un algoritmo de computación, para ser usado con una gran cantidad de información, se desea conocer la relación que existe entre el número de datos y el tiempo que emplea una computadora en entregar resultados. Para cada una de las distintas cantidades de datos, se hicieron 10 mediciones del tiempo empleado y luego se calculó el tiempo promedio, según se resumen en la tabla:

No. de datos (millones)	1.0	1.5	2.0	2.5	3.0	3.5	4.0	5.0	7.5	10
Tiempo promedio (seg.)	0.8	2.1	3.0	4.4	6.8	9.7	13.6	22.3	40.3	72.4

Ajustar los datos a un modelo multiplicativo.

Solución: El modelo propuesto es de la forma $y = \alpha x^{\lambda}\varepsilon$. Para transformarlo en uno lineal se deberá tomar logaritmos:

$$\ln y = \ln \alpha + \lambda \ln x + \ln \varepsilon.$$

Si suponemos que se cumplen las hipótesis básicas, que permiten tener un modelo de regresión lineal simple, entonces se tiene la siguiente ecuación de predicción $\ln \widehat{y} = \ln a + b \ln x$.

Por tanto, se deberá realizar una regresión lineal de $\ln y$ respecto a $\ln x$.

Tomando logaritmos en las dos variables:

$\ln x$	0.0000	0.4055	0.6931	0.9163	1.0986	1.2528	1.3863	1.6094	2.0149	2.3026
$\ln y$	−0.2231	0.4719	1.0986	1.4816	1.9169	2.2721	2.6101	3.1046	3.6964	4.2822

Efectuando la regresión lineal propuesta resulta la ecuación

$$\ln \widehat{y} = -1.866 + 1.956 \ln x.$$

A continuación, si tomamos antilogaritmos en la igualdad obtenida, da como resultado:

$$\begin{aligned} e^{\ln \widehat{y}} &= e^{-1.866+1.956 \ln x} \\ \widehat{y} &= e^{-1.866} e^{1.956 \ln x} \\ \widehat{y} &= 0.829\, x^{1.956}. \end{aligned}$$

Además, se obtiene un coeficiente de determinación muy alto, del 99 %, para la ecuación transformada.

A partir de la ecuación se podría proponer que el tiempo de ejecución del algoritmo es proporcional al cuadrado del número de datos: $y = 0.83x^2$. Para verificarlo es necesario realizar las pruebas de hipótesis sobre los parámetros. Esta tarea se deja al lector.

10.9. Ejercicios

Modelos lineales

1. Se desea estudiar la relación entre la intensidad de regadío (x) y la productividad (y) de un cierto cultivo. Se obtuvieron los siguientes resultados:

x_i	9	10	13	15	18	13
y_i	36	44	48	63	70	45

   Ajuste un modelo lineal simple y calcule el coeficiente de correlación lineal entre las variables.

2. Se realizó un experimento para medir la velocidad del sonido en el aire a diferentes temperaturas. Los resultados obtenidos se indican en la siguiente tabla:

Temperatura en °C (x)	−13	0	9	20	33	50
Velocidad en m/s (y)	322	335	337	346	352	365

   a) Estime la función de regresión lineal de y sobre x. Interprételos;
   b) Calcule el coeficiente de correlación lineal entre las variables;
   c) Pruebe si $\beta_1 = 0$;
   d) Encuentre los intervalos de estimación y de predicción cuando la temperatura es 15 °C.

3. Para determinar la relación entre el número de vendedores y las ventas anuales (en miles de dólares) que tiene una empresa, se obtuvieron los siguientes datos:

Número de vendedores	12	13	14	15	16
Ventas	20	27	33	41	58

   Se postuló que los datos se ajustan a un modelo lineal simple.

a) Determine las estimaciones de los parámetros β_0 y β_1 y escriba el modelo de regresión;

b) Grafique los puntos y la recta de ajuste;

c) Construya la tabla de análisis de la varianza y pruebe si $\beta_1 = 0$;

d) ¿Se ajustan bien los datos a la recta?

4. En un estudio para describir la relación entre la exposición al ruido y la hipertensión se realizaron las siguientes mediciones:

y	1	1	5	4	2	5	6	4	7	7
x	60	65	70	80	80	85	90	90	94	100

a) Realice un gráfico de los datos y diga si es recomendable ajustarlos a un modelo lineal simple;

b) Halle la ecuación de regresión simple que ajusta los datos;

c) Realice un análisis de varianza y pruebe la significación del ajuste;

d) Halle el coeficiente de determinación del modelo. Interprete su valor.

e) ¿Hay alguna manera de simplificar el modelo?

5. En una investigación de las propiedades de un pegamento de secado rápido se midió el tiempo que se demora en cristalizarse en función de la cantidad de pega depositada sobre una superficie de material cerámico de prueba.

Cantidad (g)	1.0	1.2	1.4	1.6	1.8	2.0
Tiempo (seg)	26.2	27.9	29.4	30.5	31.0	34.3

a) Ajuste los datos a un modelo lineal simple;

b) Realice un análisis de varianza y pruebe la significación del ajuste;

c) Construya los intervalos de confianza para los parámetros de regresión.

6. En una agencia bancaria se registró el número de depósitos realizados y el monto total de estas transacciones, en una hora de trabajo, con los siguientes resultados.

Monto total (en cientos de dólares)	10	5	7	19	11	8
Número de depósitos	16	9	3	25	7	13

a) Realice la formulación matricial del problema y ajuste los datos a un modelo lineal. Interprete los coeficientes;

b) Calcule s y obtenga un intervalo de confianza, al 95 %, para los coeficientes de regresión;

c) Evalúe r^2 e interprete su valor. Pruebe si $\rho = 0$;

d) Realice una predicción para cuando el número de depósitos es 12;

e) Obtenga la tabla ANOVA y realice la prueba F correspondiente.

7. En el mercado inmobiliario se realiza el avalúo de una propiedad para luego efectuar su venta, la diferencia constituye la ganancia del vendedor. En la tabla se dan los valores (en miles de dólares) de avalúo y precio de venta de doce propiedades en Quito.

Avalúo	**Venta**	**Avalúo**	**Venta**
46.5	59.0	67.4	108.0
43.5	56.5	70.1	95.0
52.2	65.2	74.0	84.0
62.5	74.0	57.4	106.0
85.7	109.0	103.8	154.0

a) Grafique las variables en un diagrama de dispersión;

b) Realice la formulación matricial del problema y encuentre los estimadores de los parámetros del modelo;

c) ¿Aporta el valor de avalúo x información para conocer el precio de venta y?;

d) Calcule r^2. ¿Amerita realizar una prueba de hipótesis para probar si $\rho = 0$?;

e) Obtenga un intervalo de confianza de 95 % para el precio medio de una propiedad con un avalúo de 72 mil dólares. Interprete el intervalo;

f) Encuentre un intervalo de confianza al 95 % para el precio de venta de una propiedad avaluada en 65 mil dólares. Interprete el intervalo.

8. En una entidad financiera se desea tener un método que permita realizar pronósticos de las ganancias obtenidas en base a información que pueda estar disponible de manera rápida. El gerente de crédito plantea un modelo que relaciona el número de préstamos realizados en un mes y la ganancia obtenida en el mismo período. Para tal efecto recoge la siguiente información de los 8 últimos meses:

No. préstamos	125	131	142	127	140	121	136	133
Ganancia	44	54	77	35	80	41	66	65

Los valores de las ganancias están en cientos de dólares.

a) Encuentre la ecuación de regresión lineal simple que relaciona el número de préstamos y la ganancia;

b) ¿Es satisfactorio el modelo obtenido? ¿Por qué?;

c) Realice un análisis de la varianza;

d) Haga una predicción para un mes en el que se otorgaron 123 préstamos y construya un intervalo al 95 % para tal predicción;

e) ¿Cree que podría mejorarse el modelo propuesto? ¿Cómo?

9. La siguiente tabla muestra la captura de anchoas (captura, en millones de toneladas métricas) y el precio de la harina de pescado (precio, en dólares por tonelada) para los últimos 10 años:

Año	1	2	3	4	5	6	7	8	9	10
Precio	190	160	134	129	172	239	542	245	454	410
Captura	7.23	8.53	9.82	10.26	8.96	4.45	1.78	3.3	0.8	0.5

Construya los modelos lineales que relacionen las variables $(x - y)$ e interprete los coeficientes:

a) Precio y año;

b) Captura y año;

c) Precio y captura;

Con el modelo que tenga la máxima correlación:

d) Realice la tabla ANOVA e interprétela;

e) Construya los intervalos de confianza para los coeficientes de regresión;

f) Realice la estimación de y cuando $x = 5.5$.

10. Los siguientes datos corresponden al ritmo cardiaco en reposo (Y) y el peso (X, en kg) de 6 personas.

X		Y
90		62
86		45
67		40
89		55
81		64
75		53
$\sum x_i = 488$,		$\sum y_i = 319$,
$\sum x_i^2 = 40\,092$,	$\sum x_i y_i = 26\,184$,	$\sum y_i^2 = 17\,399$.

a) Grafique los datos y examine si parece que hay una relación lineal entre las dos variables;

b) Calcule los estimadores de los parámetros de regresión;

c) Obtenga la estimación por intervalo de la media cuando $x = 88$, al nivel 95 %;

d) Obtenga el intervalo de predicción de la media cuando $x = 88$, al nivel 95 %;

e) Calcule los coeficientes de determinación y de correlación entre las dos variables.

11. En un laboratorio de prueba de automóviles se mide la distancia de frenado en relación con la velocidad que lleva el auto, dando los siguientes resultados de 20 observaciones.

$$\overline{x} = 50, \quad \sum_{i=1}^{n}(x_i - \overline{x})^2 = 1600, \quad \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = 800,$$
$$\overline{y} = 30, \quad \sum_{i=1}^{n}(y_i - \overline{y})^2 = 832.$$

a) Calcule la ecuación de ajuste para el modelo de regresión lineal simple;

b) Realice una tabla de análisis de la varianza;

c) Determine los intervalos de confianza de la estimación y de la predicción al 95 %, para el valor de la distancia de frenado, cuando la velocidad es $x_p = 60$.

12. Se realizó un estudio para determinar el efecto que tiene la temperatura (x) sobre la cantidad de gas residual generado (y) en un proceso químico. Se analizaron 12 unidades muestrales y se midieron las siguientes cantidades.

$$n = 12, \quad \overline{x} = 7.5, \quad \overline{y} = 55.1,$$
$$\sum_{i=1}^{12} x_i^2 = 4250, \quad \sum_{i=1}^{12} y_i^2 = 279\,360, \quad \sum_{i=1}^{12} x_i\, y_i = -23\,707.5.$$

a) Encuentre la ecuación de regresión lineal que explica la cantidad de gas por la temperatura del proceso;

b) Calcule el coeficiente de determinación;

c) Realice un análisis de varianza. Interprételo.

13. Una teoría financiera sostiene que hay una relación directa entre el riesgo de una inversión y el rendimiento que promete. El riesgo de una acción se mide por su valor, llamado β. En la tabla se muestran los rendimientos y valores de 12 acciones:

Rendimiento	5.4	8.9	2.3	1.5	3.7	8.2	5.3	0.5	1.3	5.9	6.8	7.2
Valor Beta	1.5	1.9	1.0	0.5	1.5	1.8	1.3	–0.5	0.5	1.8	1.9	1.9

Al ajustar un modelo de regresión a estos datos, se obtuvo las siguientes salidas:

Predictor	**Coef.**	**Error Est.**	t_{obs}
Intercepción	0.44	0.122	3.62
Pendiente	0.19	0.022	8.56

Fuente de variación	**g.l.**	SC	MC	F
Regresión	—	—	—	—
Residual	—	0.439	—	
Total	11	3.649		

a) Pruebe si los coeficientes del modelo son significativos. Escriba las pruebas utilizadas con sus valores y las regiones de rechazo;

b) Complete la tabla ANOVA e interprétela;

c) Calcule el coeficiente de determinación e interprételo.

14. Suponga que se ha ajustado una línea recta a un conjunto de 9 parejas de observaciones, dando:

$$\widehat{y} = -5 + 2x.$$

Además, se obtuvieron las siguientes desviaciones: $(x_i - \overline{x})$: -4, -3, -2, -1, 0, 1, 2, 3, 4 y la tabla de análisis de la varianza:

Fuente de variación	**g.l.**	SC	MC	F
Regresión	—	—	—	—
Residual	—	—	—	
Total	—	380		

a) Complete la tabla ANOVA;

b) Realice una prueba de adecuación del modelo.

Modelos no lineales

15. A continuación se presentan 7 mediciones de dos variables

x	0.5	1.0	1.5	2.0	2.5	3.0	3.5
y	0.6	2.7	12.2	54.6	244.7	1096.6	4914.8

Encuentre la ecuación de regresión que ajusta los datos, según un modelo exponencial y calcule el coeficiente de determinación.

16. En la siguiente tabla se encuentra el número de años para el vencimiento y el rendimiento de unos bonos.

Años para el vencimiento	1	2	5	10	15	18	23	25
Rendimiento	0.067	0.072	0.076	0.079	0.081	0.077	0.082	0.078

a) Ajuste un modelo multiplicativo a los datos;

b) Ajuste un modelo $\widehat{y} = b_0 + \dfrac{b_1}{t}$;

c) ¿Cuál de los dos modelos cree es mejor? ¿Por qué?

17. Un ingeniero está investigando la relación entre la velocidad del viento y la cantidad de electridad generada. Luego de 10 mediciones obtuvo:

Velocidad del viento	6.0	6.0	3.4	6.4	10.0	9.7	9.6	3.1	8.2	6.2
Corriente generada	1.58	1.82	1.06	1.93	2.24	2.39	2.29	0.56	2.17	1.87

a) Ajuste los datos a un modelo de tipo $y = \beta_0 + \beta_1 \left(\frac{1}{x}\right) + \varepsilon$;

b) Calcule e interprete el coeficiente de determinación del modelo transformado.

18. El gerente de una empresa desea relacionar la evolución de las ventas y el gasto publicitario, según los datos que aparecen en el cuadro:

Ventas (y)	10	15	18	23	25
Gasto (x)	19	22	41	72	98

a) Realice un ajuste de tipo multiplicativo y encuentre la calidad del ajuste;

b) Ajuste los datos mediante una función lineal simple y compare con el ajuste anterior;

c) ¿Cuál de los dos modelos es mejor? Explique.

19. Considérese los datos que se presentan a continuación

x	9	14	17	11	8	10	5	7
y	0.34	0.26	0.18	0.30	0.40	0.27	0.48	0.42

a) Grafique los datos;

b) Suponga que las variables x y y se vinculan mediante la relación $y = \dfrac{1}{\beta_0 + \beta_1 x + \varepsilon}$. Encuentre los coeficientes de regresión.

c) ¿El modelo es adecuado para los datos? ¿Por qué?

20. A continuación se presenta la evolución anual del salario mínimo vital en un país con alto índice de inflación.

Año	1	2	3	4	5	6	7	8	9	10	11	12
S.M.V.	66	95	120	120	145	190	220	320	320	400	600	600

a) Grafique los datos;

b) Ajuste los datos mediante un modelo exponencial;

c) ¿Qué se puede decir de la calidad de ajuste de los datos a la curva de regresión?;

d) Realice una estimación del valor del S.M.V. en julio de 1994;

e) Si el S.M.V. en el año 13 ascendió a 900, ¿es buena la predicción realizada? Explique.

21. En astronomía se denomina año sideral al número de años terrestres que un planeta se demora en completar una revolución alrededor del Sol y depende de la distancia entre los dos astros. En la tabla se muestra la distancia promedio y el año sideral para los planetas del Sistema Solar. Emplear los datos para determinar un modelo de regresión que relacione las dos variables,

tomando como variable dependiente al año sideral. (Para realizar la transformación adecuada refiérase a la tercera ley de Kepler).

Planeta	Distancia al Sol (r: millones de km)	Duración del año sideral (T: años terrestres)
Mercurio	58	0.241
Venus	108	0.615
Tierra	150	1.000
Marte	228	1.880
Júpiter	778	11.862
Saturno	1428	29.458
Urano	2871	84.018
Neptuno	4500	164.780
Plutón	5913	248.400

22. Los siguientes datos corresponden al precio de venta (en cientos de dólares) de un modelo de automóvil, según los años de uso

Años de uso	0	1	2	3	4	5	6
Precio	10.2	8.3	6.9	5.5	4.0	3.5	3.3

Ajuste a los datos un modelo recíproco y calcule el coeficiente de determinación.

23. Una empresa de telefonía celular ha registrado la siguiente evolución en el número de abonados a su servicio en sus primeros 8 años de operación:

Año	**Abonados**
1	32 000
2	37 500
3	41 000
4	58 000
5	107 000
6	138 000
7	175 000
8	321 500

a) Grafique los datos;

b) Se planteó que el modelo lineal simple. Realice la estimación de los coeficientes;

c) ¿Qué se puede decir de la calidad de ajuste de los datos?,

d) Efectue las pruebas de hipótesis sobre los coeficientes del modelo. Si es posible excluir alguno de ellos, ¿cómo quedaría el modelo final?;

e) Los datos sugieren una relación del tipo exponencial. Realice un ajuste de los datos al modelo propuesto;

f) Encuentre el coeficiente de determinación y compárelo con el del modelo anterior. ¿Cuál es mejor? ¿Por qué?

Capítulo 11

Regresión Múltiple

The Ballade of Multiple Regression
If you want to deal best with your questions,
Use multi-regression techniques;
A computer can do in a minute
What, otherwise done, would take weeks.
For 'predictor selection' procedures
Will pick just ones best for you
And provide the best-fitting equation
For the data you've fitted it to.
Tom Corlett, 1963

En el capítulo anterior estudiamos el caso en el cual la variable de respuesta y depende de una sola variable predictora x, estableciendo el modelo de regresión lineal simple. Pero, podría suceder que este modelo sea insuficiente y que sea necesario incorporar nuevas variables explicativas del fenómeno investigado.

Analicemos el siguiente caso: el gerente de una empresa desea incrementar las ventas, para lo cual decide realizar gastos en publicidad y medir la variación en sus ventas. Inicialmente, decide poner publicidad en televisión, pero posteriormente decide también ponerla en la radio y los periódicos.

En la primera etapa la variable de respuesta, que es el incremento en las ventas, depende de una sola variable predictora (los gastos en televisión) y para realizar un análisis es suficiente emplear un modelo de regresión lineal simple. Mas en la segunda etapa, la variable de respuesta depende de varias variables predictoras (los gastos en televisión, radio y prensa), consecuentemente para realizar un análisis ya no es suficiente la regresión lineal simple.

En general, aunque hay muchos problemas prácticos que atañen a variables predictoras únicas, es mucho más frecuente que la variable respuesta dependa de un conjunto de variables predictoras o de transformaciones de las mismas. De la estimación de tales modelos y de su calidad de ajuste nos ocuparemos en el presente capítulo.

11.1. Modelo de regresión múltiple

Definición (de modelo de regresión lineal múltiple) El modelo de regresión que liga a una variable dependiente y con k variables independientes mediante la ecuación

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{11.1}$$

se llama modelo de regresión lineal múltiple con k variables regresoras.

Los parámetros β_j, $j = 0, 1, \ldots, k$ se denominan coeficientes de regresión.

Al igual que en el caso de una sola variable, se considera que el error ε tiene esperanza igual a cero y varianza σ^2 y que los errores ε_i, correspondientes a cada observación, son no correlacionados.

Un modelo plausible para el ejemplo examinado es

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon,$$

donde y denota el incremento en las ventas, x_1 los gastos en televisión, x_2 los gastos en radio y x_3 los gastos en prensa. Los coeficientes β_0, β_1, β_2 y β_3 son parámetros que definen el modelo, ellos son desconocidos y el problema inicial es determinar estos coeficientes.

La ecuación (11.1), geométricamente representa un hiperplano en un espacio de k dimensiones. El coeficiente β_j $(j = 0, 1, \ldots, k)$ evidencia la variación unitaria de la respuesta y cuando varía x_j y las restantes variables permanecen constantes.

Mediante la técnica de regresión lineal múltiple se puede analizar una serie de modelos particulares como el polinomial de una variable

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \varepsilon,$$

o de dos o más variables; por ejemplo, el de segundo grado con dos variables es

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \varepsilon.$$

Otros modelos más complejos pueden reducirse a un modelo de regresión lineal múltiple mediante cambios de variable adecuados.

11.2. Estimación de los parámetros

Al igual que en el método de regresión lineal simple, para la estimación de los parámetros se aplica el método de mínimos cuadrados.

Suponga que disponemos de $n > k$ observaciones, y si se denota como x_{ij} al valor de la $i-$ésima observación de la variable x_j, como se observa en la siguiente tabla:

y	x_1	x_2	$\cdots$	x_k
y_1	x_{11}	x_{12}	$\cdots$	x_{1k}
y_2	x_{21}	x_{22}	$\cdots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
y_n	x_{n1}	x_{n2}	$\cdots$	x_{nk}

Si $\widehat{y}$ es la predicción de y, la ecuación de regresión queda como

$$\widehat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k, \tag{11.2}$$

donde $b_0, b_1, \ldots, b_k$ son tales que la suma de los cuadrados de las diferencias entre los valores observados de la variable respuesta y su estimación por la ecuación de regresión sea mínima.

Escribamos la igualdad (11.2) para cada una de las observaciones:

$$\begin{aligned}
\widehat{y}_1 &= b_0 + b_1x_{11} + b_2x_{12} + \cdots + b_kx_{1k} \\
\widehat{y}_2 &= b_0 + b_1x_{21} + b_2x_{22} + \cdots + b_kx_{2k} \\
\vdots &= \vdots \\
\widehat{y}_n &= b_0 + b_1x_{n1} + b_2x_{n2} + \cdots + b_kx_{nk},
\end{aligned}$$

o en forma abreviada

$$\widehat{y}_i = b_0 + \sum_{j=1}^{k} b_jx_{ij}, \quad i = 1, 2, \ldots, n.$$

Se minimizará la suma de los cuadrados de los errores

$$SCE = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n}\left[y_i - \left(b_0 + \sum_{j=1}^{k} b_jx_{ij}\right)\right]^2.$$

Derivando SCE con respecto a cada b_i, e igualando el resultado a cero se obtienen las ecuaciones

$$\begin{aligned}
\frac{\partial(SCE)}{\partial b_0} &= -2\sum_{i=1}^{n}(y_i - b_0 - \sum_{j=1}^{k} b_jx_{ij}) = 0, \\
\frac{\partial(SCE)}{\partial b_1} &= -2\sum_{i=1}^{n}x_{i1}(y_i - b_0 - \sum_{j=1}^{k} b_jx_{ij}) = 0, \\
\vdots &= \vdots \\
\frac{\partial(SCE)}{\partial b_k} &= -2\sum_{i=1}^{n}x_{ik}(y_i - b_0 - \sum_{j=1}^{k} b_jx_{ij}) = 0.
\end{aligned}$$

Luego de simplificar las igualdades, se obtiene las *ecuaciones normales* de mínimos cuadrados:

$$\begin{aligned}
nb_0 + b_1\sum_{i=1}^{n}x_{i1} + b_2\sum_{i=1}^{n}x_{i2} + \cdots + b_k\sum_{i=1}^{n}x_{ik} &= \sum_{i=1}^{n}y_i \\
b_0\sum_{i=1}^{n}x_{i1} + b_1\sum_{i=1}^{n}x_{i1}^2 + b_2\sum_{i=1}^{n}x_{i1}x_{i2} + \cdots + b_k\sum_{i=1}^{n}x_{i1}x_{ik} &= \sum_{i=1}^{n}x_{i1}y_i \\
\vdots &= \vdots \\
b_0\sum_{i=1}^{n}x_{ik} + b_1\sum_{i=1}^{n}x_{ik}x_{i1} + b_2\sum_{i=1}^{n}x_{ik}x_{i2} + \cdots + b_k\sum_{i=1}^{n}x_{ik}^2 &= \sum_{i=1}^{n}x_{ik}y_i.
\end{aligned}$$

Entonces, se dispone de un sistema de $k+1$ ecuaciones normales que involucran a los coeficientes desconocidos. Su solución permite conocer los estimadores de los parámetros del modelo lineal múltiple, aunque debe observarse que tal tarea, en general, suele ser muy laboriosa.

Formulación matricial

De manera similar a la realizada en el modelo lineal simple, el lineal múltiple es factible ponerlo en forma matricial. Por la cantidad de variables involucradas esta formulación es más fácil de manipular que la forma algebraica. El proceso es idéntico al explicado en el capítulo anterior; sin embargo, lo repetiremos de manera simplificada.

Si se ponen las matrices

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix},$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

se obtiene la ecuación matricial del modelo

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Los miembros de esta ecuación se forman de la siguiente manera:

$\mathbf{Y}$ es un vector $n \times 1$ de las observaciones.

$\mathbf{X}$ es una matriz $n \times (k+1)$ de los valores de la variable independiente, cuya primera columna tiene la particularidad de que todos sus componentes son iguales a 1.

$\boldsymbol{\beta}$ es un vector $(k+1) \times 1$ de los parámetros de la ecuación.

$\boldsymbol{\varepsilon}$ es un vector $n \times 1$ de los errores aleatorios.

Si $\mathbf{b}$ es el vector de los estimadores de los parámetros de regresión:

$$\mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix},$$

el sistema de estimación de los parámetros es $\widehat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$.

El vector de errores es igual a $\boldsymbol{\varepsilon} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\mathbf{b}$.

Por lo tanto,

$$\begin{aligned} SCE &= \boldsymbol{\varepsilon}^t\boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\mathbf{b})^t(\mathbf{Y} - \mathbf{X}\mathbf{b}) \\ &= \mathbf{Y}^t\mathbf{Y} - 2\mathbf{b}^t\mathbf{X}^t\mathbf{Y} + \mathbf{b}^t\mathbf{X}^t\mathbf{X}\mathbf{b}. \end{aligned}$$

Derivando SCE respecto a $\mathbf{b}$, e igualando el resultado a cero da

$$\mathbf{X}^t\mathbf{X}\mathbf{b} = \mathbf{X}^t\mathbf{Y}. \tag{11.3}$$

Esta expresión es la forma matricial de las ecuaciones normales de regresión antes deducidas.

Si la matriz $\mathbf{X}^t\mathbf{X}$ es inversible se obtiene el estimador $\mathbf{b}$ como la solución del sistema (11.3):

$$\mathrm{b} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y} \tag{11.4}$$

Desarrollemos la ecuación (11.3):

$$\begin{pmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik}^2 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik}y_i \end{pmatrix}.$$

Observemos la estructura especial de la matrices $\mathbf{X}^t\mathbf{Y}$ y $\mathbf{X}^t\mathbf{X}$; ésta última es simétrica de orden $k+1$.

Ejemplo. Una compañía, con objeto de programar la producción para el resto del año, necesita un pronóstico de las ventas totales. En sus registros tiene las ventas realizadas (en miles de dólares) y los gastos en publicidad e insumos para sus productos en los últimos 10 meses como se muestra en la siguiente tabla.

Ventas	Publicidad	Insumos	Ventas	Publicidad	Insumos
333	55	96	348	67	86
310	59	84	327	76	78
311	69	74	275	59	66
290	65	66	311	77	67
342	71	91	298	64	75

Encontrar la ecuación de regresión que relacione las ventas con los gastos en publicidad y en insumos.

Solución: Nombremos la variable ventas como y, los gastos en publicidad como x_1 y los gastos en insumos como x_2. El modelo de regresión es

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon.$$

La matriz $\mathbf{X}$ y el vector $\mathbf{Y}$ son

$$\mathbf{X} = \begin{pmatrix} 1 & 55 & 96 \\ 1 & 59 & 84 \\ 1 & 69 & 74 \\ 1 & 65 & 66 \\ 1 & 71 & 91 \\ 1 & 67 & 86 \\ 1 & 76 & 78 \\ 1 & 59 & 66 \\ 1 & 77 & 67 \\ 1 & 64 & 75 \end{pmatrix}, \qquad \mathbf{Y} = \begin{pmatrix} 333 \\ 310 \\ 311 \\ 290 \\ 342 \\ 348 \\ 327 \\ 275 \\ 311 \\ 298 \end{pmatrix}$$

y $\mathbf{X}^t$ es

$$\mathbf{X}^t = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 55 & 59 & 69 & 65 & 71 & 67 & 76 & 59 & 77 & 64 \\ 96 & 84 & 74 & 66 & 91 & 86 & 78 & 66 & 67 & 75 \end{pmatrix}.$$

La matriz $\mathbf{X}^t\mathbf{X}$ resulta:

$$\mathbf{X}^t\mathbf{X} = \begin{pmatrix} 10 & 662 & 783 \\ 662 & 44\,304 & 51\,636 \\ 783 & 51\,636 & 62\,335 \end{pmatrix}.$$

Su inversa $(\mathbf{X}^t\mathbf{X})^{-1}$ es

$$(\mathbf{X}^t\mathbf{X})^{-1} = \begin{pmatrix} 21.07764 & -0.18441 & -0.11200 \\ -0.18441 & 0.002266 & 0.000438 \\ -0.11200 & 0.000438 & 0.001059 \end{pmatrix}$$

y el cálculo de $\mathbf{X}^t\mathbf{Y}$ da

$$\mathbf{X}^t\mathbf{Y} = \begin{pmatrix} 3145 \\ 208\,608 \\ 248\,055 \end{pmatrix}.$$

Finalmente, multiplicando los dos últimos resultados llegamos a que $\mathbf{b}$ es

$$\mathbf{b} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 37.30 \\ 1.717 \\ 2.088 \end{pmatrix}.$$

Con todo ésto, la ecuación de regresión es

$$\widehat{y} = 37.30 + 1.717\,x_1 + 2.088\,x_2.$$

Propiedades de b

El estimador de mínimos cuadrados $\mathbf{b}$ tiene las siguientes propiedades:

1. Es insesgado para $\boldsymbol{\beta}$.
   Puesto que $\mathbf{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ y $\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{X} = \mathbf{I}$ se tiene:

$$\begin{aligned} \mathbf{E}(\mathbf{b}) &= \mathbf{E}\left[\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{Y}\right] = \mathbf{E}\left[\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\left(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\right)\right] \\ &= \mathbf{E}\left[\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} + \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\boldsymbol{\varepsilon}\right] = \boldsymbol{\beta}. \end{aligned}$$

2. La matriz de covarianza de $\mathbf{b}$ viene dada por

$$\mathrm{Cov}(\mathbf{b}) = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}.$$

   La matriz $\mathrm{Cov}(\mathbf{b})$ es simétrica; además, el valor de σ^2 suele ser desconocido, debiendo ser estimado.

Estimación de σ^2

Consideremos ahora la suma de los cuadrados de los errores

$$SCE = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \boldsymbol{\varepsilon}^t\boldsymbol{\varepsilon},$$

que después de sustituir el valor de ε se llega a

$$SCE = \mathbf{Y}^t\mathbf{Y} - \mathbf{b}^t\mathbf{X}^t\mathbf{Y},$$

que tiene $n-k-1$ grados de libertad.

Así, el error cuadrático medio, que es un estimador insesgado de σ^2, se calcula por

$$MCE = s^2 = \frac{SCE}{n-k-1}.$$

Ejemplo. (Continuación) Estimar la varianza del error σ^2 y encontrar la matriz de varianza-covarianza de $\mathbf{b}$.

Solución: Utilizando el vector de estimación de los parámetros se obtiene

$$\mathbf{b}^t\mathbf{X}^t\mathbf{Y} = \begin{pmatrix} 37.30 & 1.717 & 2.088 \end{pmatrix} \begin{pmatrix} 3145 \\ 208\,608 \\ 248\,055 \end{pmatrix} = 993\,566.63$$

$$\mathbf{Y}^t\mathbf{Y} = \sum_{i=1}^{10} y_i^2 = 993\,957.$$

Por lo tanto, SCE es

$$SCE = \mathbf{Y}^t\mathbf{Y} - \mathbf{b}^t\mathbf{X}^t\mathbf{Y} = 993\,957 - 993\,566.63 = 390.37.$$

Como $n-k-1 = 10-2-1 = 7$, el error cuadrado medio es

$$s^2 = MCE = \frac{390.37}{7} = 55.7666.$$

La matriz de covarianza $\text{Cov}(\mathbf{b}) = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$ se estima por

$$\begin{aligned} \widehat{\text{Cov}(\mathbf{b})} &= s^2(\mathbf{X}^t\mathbf{X})^{-1} = 55.7666 \begin{pmatrix} 21.07764 & -0.18441 & -0.11200 \\ -0.18441 & 0.002266 & 0.000438 \\ -0.11200 & 0.000438 & 0.001059 \end{pmatrix} \\ &= \begin{pmatrix} 1175.43 & 10.248 & 6.246 \\ 10.248 & 0.126 & 0.024 \\ 6.246 & 0.024 & 0.059 \end{pmatrix}. \end{aligned}$$

11.2.1. Coeficientes de regresión estandarizados

Cuando se realiza un modelo de regresión múltiple, generalmente, las unidades de medición no son las mismas para la variable dependiente y para las variables independientes, de manera que los coeficientes de regresión no se pueden comparar directamente. Para superar esta dificultad, se emplean los *coeficientes de regresión estandarizados* **beta**.

> Las unidades de medición de todas las variables se transforman a unidades de desviación estándar, dividiendo cada variable por su desviación estándar.

Por ejemplo, en la ecuación de regresión

$$\widehat{y} = b_0 + b_1x_1 + b_2x_2,$$

se tiene

$$\frac{\widehat{y}}{s_y} = \frac{b_0}{s_y} + \left(b_1 \frac{s_{x1}}{s_y}\right) \frac{x_1}{s_{x1}} + \left(b_2 \frac{s_{x2}}{s_y}\right) \frac{x_2}{s_{x2}}.$$

Los coeficientes

$$\mathbf{beta}_i = b_i \frac{s_{x_i}}{s_y}$$

son los coeficientes de regresión parcial estándar y su interpretación es la siguiente: si hay una variación de una desviación estándar en x_i, habrá una desviación de $\mathbf{beta}_i$ desviaciones estándar en y.

Ejemplo. (Continuación) Encontrar los coeficientes de regresión estandarizados de los datos del ejemplo.

Solución: Se tiene que

$$b_0 = 37.30, \quad b_1 = 1.717, \quad b_2 = 2.088.$$

Las desviaciones estándar de las variables son

$$s_y = 23.22, \quad s_{x1} = 7.30, \quad s_{x2} = 10.68.$$

Los coeficientes beta son:

$$\begin{aligned} \mathbf{beta}_1 &= b_1 \frac{s_{x1}}{s_y} = 1.717 \frac{7.30}{23.22} = 0.5398, \\ \mathbf{beta}_2 &= b_2 \frac{s_{x2}}{s_y} = 2.088 \frac{10.68}{23.22} = 0.9604. \end{aligned}$$

11.3. Intervalos de confianza

En la regresión múltiple se pueden formular estimaciones por intervalo para los valores de los coeficientes de regresión, para los valores estimados y para nuevas predicciones.

11.3.1. Intervalos de confianza para β

Puesto que el vector de coeficientes $\boldsymbol{\beta}$ es desconocido, lo consideramos como una variable aleatoria multivariante, normalmente distribuida con media $\mathbf{b}$ y matriz de covarianza $\sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$, por lo que cada uno de los estadísticos

$$\frac{b_j - \beta_j}{s\sqrt{C_{jj}}}, \qquad j = 0, 1, \ldots, k$$

sigue una ley t con $(n-k-1)$ grados de libertad y donde C_{jj} es el $j-$ésimo elemento de la diagonal de la matriz $(\mathbf{X}^t\mathbf{X})^{-1}$.

Un intervalo de confianza al $100(1-\alpha)\,\%$ para el coeficiente de regresión β_j $(j = 0, 1, \ldots, k)$, es

$$\left(b_j - t_{\alpha/2}(n-k-1)\, s\sqrt{C_{jj}}; b_j + t_{\alpha/2}(n-k-1) s\, \sqrt{C_{jj}}\right).$$

Ejemplo. (Continuación) Elaborar un intervalo de confianza al 95 % para los parámetros β_0, β_1 y β_2, estimados anteriormente.

Solución: El estimador de σ^2 es $s^2 = 55.7666$ y $t_{0.025}(7) = 2.365$; entonces, se tienen los siguientes intervalos:

1. Para β_0:

$$\left(b_0 - t_{\alpha/2}(n-k-1)\, s\sqrt{C_{00}}; b_0 + t_{\alpha/2}(n-k-1)\, s\sqrt{C_{00}}\right)$$
$$\left(37.30 - 2.365 \times 7.47 \times \sqrt{21.077}; 37.30 + 2.365 \times 7.47 \times \sqrt{21.077}\right)$$
$$(-43.80; 118.40)$$

2. Para β_1:

$$\left(b_1 - t_{\alpha/2}(n-k-1)\, s\sqrt{C_{11}}; b_1 + t_{\alpha/2}(n-k-1)\, s\sqrt{C_{11}}\right)$$
$$\left(1.717 - 2.365 \times 7.47 \times \sqrt{0.0022}; 1.717 + 2.365 \times 7.47 \times \sqrt{0.0022}\right)$$
$$(0.888; 2.546)\,.$$

3. Para β_2:

$$\left(b_2 - t_{\alpha/2}(n-k-1)s\sqrt{C_{22}}; b_2 + t_{\alpha/2}(n-k-1)s\sqrt{C_{22}}\right)$$
$$\left(2.088 - 2.365 \times 7.47 \times \sqrt{0.001}; 2.088 + 2.365 \times 7.47 \times \sqrt{0.001}\right)$$
$$(1.515; 2.647).$$

11.3.2. Intervalo de confianza para la estimación, E(Y)

Si se desea conocer el intervalo de confianza para la respuesta media de un punto en particular x_{p1}, x_{p2}, ..., x_{pk}, definimos el vector

$$\mathbf{x}_p = \begin{pmatrix} 1 \\ x_{p1} \\ x_{p2} \\ \vdots \\ x_{pk} \end{pmatrix}.$$

La respuesta en este punto es

$$\widehat{y}_p = \mathbf{x}_p^t \mathbf{b}.$$

Este estimador es insesgado (es decir, $\mathbf{E}\left(\widehat{Y}_p\right) = \mathbf{x}_p^t \boldsymbol{\beta}$) y su varianza es

$$\text{Var}\left(\widehat{Y}_p\right) = \sigma^2 \mathbf{x}_p^t (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_p.$$

Un intervalo de confianza al $100(1-\alpha)\,\%$ para la estimación, $\mathbf{E}(Y_p)$, en el punto $\mathbf{x}_p$ es

$$\left(\widehat{y}_p - t_{\alpha/2}(n-k-1)\, s\sqrt{\mathbf{x}_p^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_p}; \widehat{y}_p + t_{\alpha/2}(n-k-1)\, s\sqrt{\mathbf{x}_p^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_p}\right).$$

11.3.3. Intervalo de confianza para la predicción, $\mathbf{y}_p$

Un modelo de regresión se aplica en la realización de pronósticos correspondientes a valores particulares de las variables independientes, $\mathbf{x}_p$. La respuesta en este punto es $\widehat{y}_p = \mathbf{x}_p^t\mathbf{b}$.

El intervalo de confianza de nivel $100(1-\alpha)\,\%$ para la predicción, y_p, es

$$\left(\widehat{y}_p - t_{\alpha/2}(n-k-1)\, s\sqrt{1+\mathbf{x}_p^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_p}; \widehat{y}_p + t_{\alpha/2}(n-k-1)\, s\sqrt{1+\mathbf{x}_p^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_p}\right).$$

Ejemplo. (Continuación) Elaborar los intervalos de confianza al 95 % para la respuesta media y para la predicción, cuando los gastos en publicidad son de 72 mil y en insumos 90 mil dólares.

Solución: El punto sobre el cual deseamos realizar la predicción es

$$\mathbf{x}_p = \begin{pmatrix} 1 \\ 72 \\ 90 \end{pmatrix}.$$

La respuesta en el punto considerado es $\widehat{y}_p = \mathbf{x}_p^t\mathbf{b} = 348.844$.

El intervalo de confianza para la estimación resulta:

$$\left(\widehat{y}_p - t_{\alpha/2}(n-k-1)\, s\sqrt{\mathbf{x}_p^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_p}; \widehat{y}_p + t_{\alpha/2}(n-k-1)\, s\sqrt{\mathbf{x}_p^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_p}\right)$$

$$(348.8 - 2.365 \times 7.47 \times \sqrt{0.381}; 348.8 + 2.365 \times 7.47 \times \sqrt{0.381})$$

$$(337.9; 359.7).$$

El intervalo de confianza para la predicción es

$$\left(\widehat{y}_p - t_{\alpha/2}(n-k-1)\, s\sqrt{1+\mathbf{x}_p^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_p}; \widehat{y}_p + t_{\alpha/2}(n-k-1)\, s\sqrt{1+\mathbf{x}_p^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_p}\right)$$

$$(348.8 - 2.365 \times 7.47 \times \sqrt{1.381}; 348.8 + 2.365 \times 7.47 \times \sqrt{1.381})$$

$$(328.0; 369.6).$$

11.4. Pruebas de hipótesis

Las pruebas de hipótesis en la regresión lineal múltiple se emplean de dos maneras: para determinar la significación de la regresión, es decir, si globalmente las variables aportan información al modelo; y, para realizar pruebas sobre los valores de los coeficientes individuales para examinar si una variable particular es significativa en el modelo y merece ser incluida en la ecuación.

11.4.1. Prueba de hipótesis para la significación del modelo

La prueba global del modelo se emplea para conocer si existe relación lineal entre la variable dependiente y y el conjunto de las variables independientes x_1, x_2, ..., x_k.

Previamente, descompongamos la suma total de los cuadrados SC_{yy} en dos sumas, una debida a la regresión, SCR, y otra debida al error, SCE:

$$SC_{yy} = SCE + SCR,$$

donde

$$\begin{aligned} SCE &= \mathbf{Y}^t\mathbf{Y} - \mathbf{b}^t\mathbf{X}^t\mathbf{Y}, \\ SCR &= \mathbf{b}^t\mathbf{X}^t\mathbf{Y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}, \\ SC_{yy} &= \mathbf{Y}^t\mathbf{Y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}. \end{aligned}$$

que nos sirven para realizar de manera ordenada todos los pasos involucrados en la prueba de hipótesis, con el empleo de una tabla de análisis de la varianza.

En el siguiente cuadro se resume los elementos de un análisis de varianza para una regresión múltiple:

Tabla de Análisis de la Varianza				
Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F
Regresión	SCR	k	MCR	$F_{obs} = \dfrac{MCR}{MCE}$
Error o residual	SCE	$n-k-1$	$s^2 = MCE$	
Total	SC_{yy}	$n-1$		

Entonces, se plantea la siguiente prueba de hipótesis:

1. *Hipótesis Nula.* H_0: $\beta_1 = \cdots = \beta_k = 0$.
2. *Hipótesis Alternativa.* H_1: $\beta_k \neq 0$ para al menos una k.
3. *Estadístico de Prueba.* $F_{obs} = \dfrac{MCR}{MCE}$.
4. *Región de Rechazo.* Se rechaza H_0 si $F_{obs} > F_\alpha(k, n-k-1)$.

El rechazo de H_0 significa que al menos una de las variables independientes x_j contribuye significativamente al modelo lineal propuesto.

Ejemplo. (Continuación) Realizar un análisis de adecuación del modelo, a un nivel $\alpha = 0.05$.

Solución: Anteriormente habíamos calculado dos componentes de las sumas de cuadrados:

$$\mathbf{Y}^t\mathbf{Y} = 993\,957, \quad \mathbf{b}^t\mathbf{X}^t\mathbf{Y} = 993\,566.63.$$

Calculemos el tercero:

$$\frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = \frac{(3145)^2}{10} = 989\,102.5.$$

También, se calculó que $SCE = 390.36$. Las restantes sumas de cuadrados son

$$\begin{aligned} SC_{yy} &= 993\,957 - 989\,102.5 = 4854.5, \\ SCR &= 993\,566.63 - 989\,102.5 = 4464.13. \end{aligned}$$

La tabla de análisis de la varianza queda así:

Tabla de análisis de la varianza				
Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F
Regresión	4464.13	2	2232.07	40.02
Error o residual	390.37	7	55.76	
Total	4854.50	9		

La prueba de hipótesis de adecuación del modelo es:

1. *Hipótesis Nula.* H_0: $\beta_1 = \beta_2 = 0$.
2. *Hipótesis Alternativa.* H_1: $\beta_1 \neq 0$ o $\beta_2 \neq 0$.
3. *Estadístico de Prueba.* $F_{obs} = 40.02$.
4. *Región de Rechazo.* El valor $F_{0.05}(2,7) = 4.74$ define la región crítica $F_{obs} > 4.74$.
5. *Decisión.* Como $40.02 > 4.74$, resulta que se rechaza la hipótesis nula y se concluye que las ventas están relacionadas con los gastos en publicidad y en insumos. ◀

Con esta prueba estadística únicamente se concluye la validez del modelo, en forma global; ella no indica la importancia relativa de cada una de las variables predictoras sobre la variable respuesta.

11.4.2. Pruebas acerca de los parámetros β individuales

En el modelo pueden estar incluidas variables que son redundantes o que no aportan significativamente a la respuesta, eso se puede comprobar realizando una prueba sobre los parámetros individuales de regresión. Entonces, se tiene la siguiente prueba de hipótesis para un parámetro β_i fijo.

1. *Hipótesis Nula.* H_0: $\beta_i = 0$.
2. *Hipótesis Alternativa.* H_1: $\beta_i \neq 0$.
3. *Estadístico de Prueba.* $t_{obs} = \dfrac{b_i}{s\sqrt{C_{ii}}}$.
4. *Región de Rechazo.* Se rechaza H_0 si $t_{obs} < -t_{\alpha/2}(n-k-1)$ o $t_{obs} > t_{\alpha/2}(n-k-1)$.

Ejemplo. (Continuación) En el modelo planteado, realizar la prueba para los parámetros β_i, $i = 0, 1, 2$, al nivel del 95 %.

Solución: Se tienen las siguientes pruebas:

a) Para el coeficiente β_0:

1. *Hipótesis Nula.* H_0: $\beta_0 = 0$.
2. *Hipótesis Alternativa.* H_1: $\beta_0 \neq 0$.
3. *Estadístico de Prueba.* $t_{obs} = \dfrac{b_0}{s\sqrt{C_{00}}} = \dfrac{37.30}{7.47\sqrt{21.077}} = 1.088$.
4. *Región de Rechazo.* Como $t_{0.025}(7) = 2.365$, la región es $|t_{obs}| > 2.365$.
5. *Decisión.* No se rechaza H_0; por tanto, la constante podría excluirse de la regresión.

b) Para β_1:

1. *Hipótesis Nula.* H_0: $\beta_1 = 0$.
2. *Hipótesis Alternativa.* H_1: $\beta_1 \neq 0$.
3. *Estadístico de Prueba.* $t_{obs} = \dfrac{b_1}{s\sqrt{C_{11}}} = \dfrac{1.717}{7.47\sqrt{0.0022}} = 4.9$.

4. *Región de Rechazo.* $|t_{obs}| > 2.365$.
5. *Decisión.* Como $4.9 > 2.365$, se rechaza H_0; entonces, la variable «gastos en publicidad» aporta información al modelo.

c) Para β_2:

1. *Hipótesis Nula.* H_0: $\beta_2 = 0$.
2. *Hipótesis Alternativa.* H_1: $\beta_2 \neq 0$.
3. *Estadístico de Prueba.* $t_{obs} = \dfrac{b_2}{s\sqrt{C_{22}}} = 8.593$.
4. *Región de Rechazo.* $|t_{obs}| > 2.365$.
5. *Decisión.* Dado que $8.593 > 2.365$, se rechaza H_0; por tanto, la variable «gastos en insumos» aporta información al modelo.

El modelo final podría ser uno en el que no se considere el término independiente: $y = \beta_1 x_1 + \beta_2 x_2$. Se recomienda al lector que recalcule tal modelo.

11.5. Coeficientes de determinación y de correlación parcial

El coeficiente de determinación se emplea como una medida de la adecuación del modelo, que informa sobre la fuerza de la relación existente entre las variables independientes y la dependiente.

11.5.1. Coeficientes de determinación múltiple

El coeficiente de determinación múltiple se define como

$$R^2 = \frac{SCR}{SC_{yy}} = 1 - \frac{SCE}{SC_{yy}}.$$

El rango de variación de R^2 es $0 \leq R^2 \leq 1$. Un valor alto de R^2, cercano a 1, significa que el modelo de regresión es bueno y si R^2 es cercano a 0, el ajuste es malo. Si $R^2 = 0$ quiere decir que falta por completo el ajuste del modelo a los datos y si $R^2 = 1$ se tiene un ajuste perfecto.

Puesto que R^2 tiende a sobrestimar el valor de la correlación entre las variables involucradas, se emplea el coeficiente de determinación ajustado, R_a^2, que está diseñado para compensar el sesgo optimista de R^2.

El coeficiente de determinación ajustado, se define por

$$R_a^2 = R^2 - \frac{k(1-R^2)}{n-k-1}.$$

Equivalentemente, R_a^2 se calcula mediante

$$R_a^2 = 1 - \frac{SCE/(n-k-1)}{SC_{yy}/(n-1)}.$$

El rango de variación de R_a^2 es $0 \leq R_a^2 \leq 1$ y su interpretación es la misma que la del coeficiente de determinación múltiple R^2.

Ejemplo. (Continuación) Calcular R^2 y R_a^2 para el modelo planteado.

Solución: Se tiene que

$$R^2 = \frac{SCR}{SC_{yy}} = \frac{4464.13}{4854.5} = 0.9196.$$

Como el valor de R^2 es alto, se concluye que el modelo se adecua a los datos y que el 91.96 % de la variabilidad en las ventas se explica mediante las variables «gastos en publicidad» y «gastos en insumos».

$$\begin{aligned} R_a^2 &= R^2 - \frac{k(1-R^2)}{n-k-1} \\ &= 0.9196 - \frac{2(1-0.9196)}{10-2-1} = 0.8966. \end{aligned}$$

En cambio, si interpretamos R_a^2, podemos decir que la calidad del ajuste es de un 90 %. Para mejorar el modelo se podría incluir una tercera variable explicativa y comprobar si ella es o no significativa.

11.5.2. Correlación parcial

El coeficiente de correlación parcial mide la correlación entre dos variables, manteniendo las demás con valores constantes; es decir, dado un conjunto x_1, x_2, ..., x_k, el coeficiente de correlación parcial entre dos cualesquiera de ellas, x_i y x_j, es una medida (adimensional) de su relación lineal, cuando se elimina de ambas los efectos debidos al resto de las variables.

Por ejemplo, con k regresores, el coeficiente de correlación parcial entre x_1 y x_2, que se denota $r_{12,34\ldots k}$, se define como el coeficiente de correlación lineal entre x_1 y x_2 cuando se elimina de ambas variables el efecto de los otros $k-2$ regresores. Se calcula obteniendo el coeficiente de correlación en la regresión

$$e_{1,34\ldots k} = \widehat{\beta} e_{2,34\ldots k} + u; \qquad (r_{12,34\ldots k}),$$

donde $e_{1,34\ldots k}$ y $e_{2,34\ldots k}$ son los residuos de la regresión múltiple de x_1 y x_2 respecto a las variables de control x_3, ..., x_k.

Si tuviéramos el modelo $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, el coeficiente de correlación parcial de y con x_1, cuando x_2 permanece constante, se define por

$$r_{y1,2} = \frac{r_{y1} - r_{y2} r_{12}}{\sqrt{(1-r_{y2}^2)(1-r_{12}^2)}},$$

donde r_{y1}, r_{y2} y r_{12} son los coeficientes de correlación de Pearson de y con x_1, de y con x_2 y de x_1 con x_2, respectivamente. De manera análoga, se tiene

$$r_{12,y} = \frac{r_{12} - r_{y1} r_{y2}}{\sqrt{(1-r_{y1}^2)(1-r_{y2}^2)}},$$

el coeficiente de correlación parcial entre x_1 y x_2.

Por lo complicado que puede resultar el cálculo de las correlaciones parciales, sus valores se obtienen mediante programas estadísticos.

Ejemplo. (Continuación) Calcular los coeficientes de correlación parcial para el modelo planteado.

Solución: Tenemos los coeficientes de correlación de Pearson entre las variables:

$$r_{12} = -0.283, \quad r_{y1} = 0.268, \quad r_{y2} = 0.807.$$

Entonces,

$$r_{y1,2} = \frac{r_{y1} - r_{y2}r_{12}}{\sqrt{(1-r_{y2}^2)(1-r_{12}^2)}} = \frac{0.268 - 0.807 \times (-0.283)}{\sqrt{(1-(0.807)^2)\,(1-(-0.283)^2)}} = 0.876,$$

$$r_{y2,1} = \frac{r_{y2} - r_{y1}r_{12}}{\sqrt{(1-r_{y1}^2)(1-r_{12}^2)}} = \frac{0.807 - 0.268 \times (-0.283)}{\sqrt{(1-(0.268)^2)\,(1-(-0.283)^2)}} = 0.955,$$

$$r_{12,y} = \frac{r_{12} - r_{y1}r_{y2}}{\sqrt{(1-r_{y1}^2)(1-r_{y2}^2)}} = \frac{-0.283 - 0.268 \times 0.807}{\sqrt{(1-(0.268)^2)\,(1-(0.807)^2)}} = -0.878.$$

11.6. Regresión polinomial

El primer paso para escoger un modelo que describa los datos, es la realización de un gráfico de dispersión de las observaciones. La relación sugerida por los datos es la que permite escoger un modelo que los describa adecuadamente.

Cuando los datos presentan un esquema de comportamiento curvilíneo puede ser necesario proponer un modelo de tipo polinomial para los datos. Así lo observamos en la Figura 11.1.

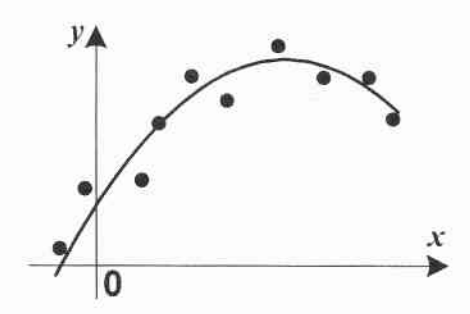


Figura 11.1: Modelo de segundo grado para los datos.

Veamos, con un ejemplo, cómo se puede transformar un modelo polinomial en uno de regresión múltiple.

Supongamos que escogemos un modelo de segundo grado en una variable:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon,$$

si hacemos las sustituciones de variables $x_1 = x$ y $x_2 = x^2$, la ecuación de regresión queda como

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

que es un modelo de regresión múltiple en dos variables. En este momento estamos en posibilidad de aplicar la teoría anteriormente descrita.

En general, si se tiene un modelo polinomial con una variable explicativa

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \varepsilon,$$

se lo convierte en uno de regresión múltiple mediante la transformación $x_i = x^i$, quedando:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon.$$

Otros modelos polinomiales que incluyen más de una variable, que pueden transformarse a uno lineal múltiple, son los polinomios en varias variables, como el de segundo grado en dos variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon.$$

Cuando se ajusta un modelo polinomial es preciso escoger el polinomio de menor grado posible. consecuentemente se deberán realizar reiteradas pruebas de hipótesis, en las que se fijarán aquellas variables que se han de incluir y excluir en el modelo final.

Ejemplo. Consideremos los datos que relacionan el número de páginas de un folleto y el costo de los insumos utilizados en la impresión de 100 ejemplares.

No. de págs.	Costo	No. de págs.	Costo
90	204	50	130
80	170	40	126
75	165	35	124
70	155	30	121
65	148	25	100
60	140	20	98

a) Ajustar un polinomio de segundo orden a los datos;

b) Probar la significación de la regresión;

c) Probar la hipótesis de que $\beta_1 = 0$.

Solución:

a) Como se observa en el gráfico de los datos, éstos podrían ajustarse a un modelo de segundo grado de la variable independiente, entonces planteamos el modelo $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$.

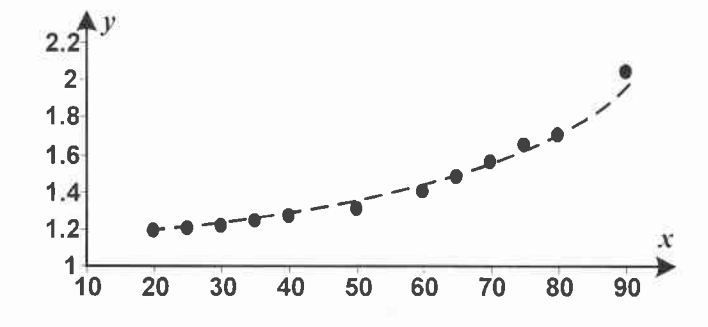


Figura 11.2: Relación entre el número de páginas y el costo de una publicación.

La matriz $\mathbf{X}$, el vector $\mathbf{Y}$ y el vector $\boldsymbol{\beta}$ son:

$$\mathbf{X} = \begin{pmatrix} 1 & 90 & 8100 \\ 1 & 80 & 6400 \\ 1 & 75 & 5625 \\ 1 & 70 & 4900 \\ 1 & 65 & 4225 \\ 1 & 60 & 3600 \\ 1 & 50 & 2500 \\ 1 & 40 & 1600 \\ 1 & 35 & 1225 \\ 1 & 30 & 900 \\ 1 & 25 & 625 \\ 1 & 20 & 400 \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} 204 \\ 170 \\ 165 \\ 155 \\ 148 \\ 140 \\ 130 \\ 126 \\ 124 \\ 121 \\ 100 \\ 98 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

Resolviendo las ecuaciones normales $\mathbf{X}^t\mathbf{X}\mathbf{b} = \mathbf{X}^t\mathbf{Y}$ se obtiene el modelo estimado

$$\widehat{y} = 1374.4 - 12.373x + 0.218x^2.$$

b) Los resultados se resumen en la siguiente tabla:

Tabla de análisis de la varianza				
Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrado medio	F
Regresión	2	6622.24	3311.12	6.91
Residual	9	4314.68	479.41	
Total	11	10 936.92		

Puesto que $F_{obs} = 6.91$, que si se compara con el valor de la tabla correspondiente a $F_{0.05}(2,9) = 4.26$, resulta que $F_{obs} > F_{0.05}(2,9)$. El resultado es significativo al nivel del 5 %; es decir, al menos uno de los parámetros β_1 o β_2 es distinto de cero. Además, $R^2 = 0.605$, que en este caso indica que la calidad del ajuste es buena, pero no lo suficiente.

c) Veamos si es posible la eliminación del término de primer grado de la ecuación.

Se tiene la prueba de hipótesis:

1. *Hipótesis Nula.* H_0: $\beta_1 = 0$.
2. *Hipótesis Alternativa.* H_1: $\beta_1 \neq 0$.
3. *Estadístico de Prueba.* $t_{obs} = -2.318$.
4. *Región de Rechazo.* Tomemos $\alpha = 0.05$. Como $t_{0.025}(9) = 2.262$, se tiene la región de rechazo $|t_{obs}| > 2.262$.
5. *Decisión.* Como $|-2.318| > 2.262$, no se puede eliminar el término de primer grado.

Se sugiere que el lector realice los análisis con el empleo de un paquete estadístico.

11.7. Regresión con variables cualitativas

En los modelos tratados se empleó variables independientes de naturaleza cuantitativa; es decir, que se expresan numéricamente y son el resultado de mediciones instrumentales. Pero si se desea incorporar

en el modelo una variable cualitativa, es necesario introducir *variables indicadoras* (o *ficticias*), que permiten diferenciar los distintos niveles que toma tal variable; por ejemplo, una variable X que indique la estación del año puede ser definida como

$$X = \begin{cases} 0, & \text{si es invierno;} \\ 1, & \text{si es verano.} \end{cases}$$

En general, una variable cualitativa con t niveles se representa mediante $t-1$ variables indicadoras, a las cuales se les asignan valores de 0 y 1.

Ejemplo. En un estudio para determinar la relación entre el peso y el origen de los automóviles y su consumo de combustible se escogió una muestra de 10 carros, con los siguientes resultados:

Consumo (l/100 km)	8	16	6	7	7	9	11	12	18	20
Peso (kg)	739	1187	655	729	888	797	963	802	1551	1650
Origen	Japón	USA	Japón	Japón	Japón	Japón	USA	USA	USA	USA

Determinar la ecuación de regresión que ajusta los datos y probar su significación.

Solución:

a) Se va a ajustar el modelo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

donde la variable peso es cuantitativa (x_1) y la variable origen es cualitativa (x_2), con dos niveles: USA y Japón, que la codificaremos de la siguiente manera:

$$x_2 = \begin{cases} 0, & \text{si el origen es Japón;} \\ 1, & \text{si el origen es USA.} \end{cases}$$

Con ésto la matriz $\mathbf{X}$ y el vector $\mathbf{Y}$ quedan:

$$\mathbf{X} = \begin{pmatrix} 1 & 739 & 0 \\ 1 & 1187 & 1 \\ 1 & 655 & 0 \\ 1 & 729 & 0 \\ 1 & 888 & 0 \\ 1 & 797 & 0 \\ 1 & 963 & 1 \\ 1 & 802 & 1 \\ 1 & 1551 & 1 \\ 1 & 1650 & 1 \end{pmatrix}, \qquad \mathbf{Y} = \begin{pmatrix} 8 \\ 16 \\ 6 \\ 7 \\ 7 \\ 9 \\ 11 \\ 12 \\ 18 \\ 20 \end{pmatrix}.$$

Y el modelo estimado es

$$\widehat{y} = -0.036 + 0.01x_1 + 3.421x_2.$$

El significado del término correspondiente al origen del automóvil es el siguiente: para dos autos, de igual peso, uno de origen americano y otro de origen japonés, el americano consume, en promedio, 3.42 litros más que el japonés, al recorrer 100 km.

b) Veamos la tabla de análisis de la varianza y las estimaciones de los parámetros:

Tabla de análisis de la varianza				
Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrado medio	F
Regresión	2	213.95	106.975	71.66
Error	7	10.45	1.493	
Total	9	22.40		

Por el alto valor de $F_{obs} = 71.66$, se deduce que al menos una de las dos variables consideradas sirve para explicar el consumo de combustible de los carros.

Término	Valor	t_{obs}
Constante b_0	-0.036	-0.027
Peso b_1	0.01	6.012
Origen b_2	3.242	3.153

Si comparamos los valores de t_{obs} con $t_{0.025}(7) = 2.365$, deducimos que los términos correspondientes al peso y origen son distintos de cero, mientras que el término constante se puede considerar nulo.

Para terminar, examinemos los valores ajustados, comparándolos con los datos originales y el error respectivo:

y_i	$\widehat{y}_i$	e_i
8	7.18	0.82
16	14.97	1.03
6	6.36	-0.36
7	7.08	-0.08
7	8.63	-1.63
9	7.75	1.25
11	12.79	-1.79
12	11.22	0.78
18	18.53	-0.53
20	19.49	0.51

11.8. Problemas en la regresión múltiple

En la estimación de los parámetros en un modelo lineal múltiple se presentan varios problemas que se deben tener en cuenta al realizar un ajuste de los datos; ellos son la multicolinealidad, la presencia de valores extremos, la autocorrelación y la no normalidad de los errores.

11.8.1. Análisis de los errores

Una vez construido el modelo de regresión se deberá comprobar si las hipótesis de linealidad, de normalidad y de independencia se cumplen.

La matriz de covarianzas de los errores es

$$\text{Cov}(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{V}),$$

donde $\mathbf{V} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$; así,

$$\text{Var}(e_i) = \sigma^2(1 - v_{ii}),$$

donde el término v_{ii} mide la distancia entre el punto $\mathbf{x}_i$ y la media $\overline{\mathbf{x}}$.

Para comparar los residuos suele ser más cómodo cambiarlos de escala, estandarizándolos o estudentizándolos.

Los residuos estandarizados se definen por

$$r_i = \frac{e_i}{s\sqrt{1 - v_{ii}}}$$

y los residuos estudentizados mediante

$$t_i = \frac{e_i}{s_{(i)}\sqrt{1 - v_{ii}}},$$

que siguen una ley t con $n-k-2$ grados de libertad; donde $s_{(i)}$ son los residuos de la regresión cuando se excluye la $i-$ésima observación.

Análisis gráfico

Una vez que se han construido los residuos (e_i, r_i o t_i) es cómodo realizar gráficos como los siguientes:

1. Histogramas o gráficos de probabilidad normal.
2. Gráficos de los residuos respecto a los valores estimados, $e_i = f(\widehat{y}_i)$.
3. Gráficos de los residuos respecto a las variables explicativas, $e_i = f(x_{ki})$.

Error de especificación

Se comete error de especificación cuando establecemos una dependencia errónea de la respuesta en función de las variables explicativas: omitimos variables importantes, introducimos variables innecesarias o suponemos una relación lineal cuando la dependencia es no lineal.

La especificación incorrecta del modelo conduce a que los residuos tengan esperanza no nula y que los estimadores obtenidos sean sesgados.

11.8.2. No-normalidad de los residuos

La suposición que los residuos ε están normalmente distribuidos no es necesaria para la estimación de los parámetros de regresión ni para la partición de la variabilidad total. La normalidad es necesaria para la construcción de las pruebas de hipótesis e intervalos de confianza sobre los parámetros.

El impacto de la no normalidad en los mínimos cuadrados depende del grado de desviación de la normalidad y de la aplicación específica.

Los estimadores de los parámetros serán insesgados, pero sus intervalos de confianza y las pruebas de hipótesis serán incorrectas. Sin embargo, la prueba F es razonablemente robusta contra la no normalidad.

Para detectar la normalidad de los errores es conveniente fijarse en los coeficientes de asimetría y de curtosis. Además, se pueden realizar gráficos Q-Q o P-P de bondad de ajuste a la ley normal.

La transformación de la variable dependiente a una forma que sea más cercana a la normal es un recurso muy empleado. Estas transformaciones suelen ser sugeridas por los gráficos de los residuos. También, se puede utilizar el método de Box-Cox de transformación potencial.

En muchos casos, la desviación de la normalidad se debe a la presencia de valores atípicos, en cuyo caso es conveniente examinar la influencia de los mismos.

Recientemente se han desarrollado modelos que consideran que los errores están distribuidos según una ley t, de un número de grados de libertad desconocido, como una generalización de la hipótesis de normalidad.

11.8.3. Presencia de puntos inusuales

Generalmente, se espera que los datos correspondientes a las observaciones se encuentren distribuidos en una región más o menos cercana, pero puede suceder que una o varias observaciones estén alejadas del resto de los datos. Estas observaciones pueden influir mucho en el modelo final.

Este aspecto es muy importante: podemos disponer de 100 observaciones y, sin embargo, construir con ellos un modelo cuyas propiedades son debidas únicamente a dos puntos. Conocer si este tipo de puntos influye perjudicialmente en el modelo permite mejorarlo.

Identificación

La primera forma para determinar si un valor es atípico es mediante los residuos estudentizados. Se compara con los valores de t_i con los valores críticos de una ley t con $n-k-2$ grados de libertad.

Otra forma de conocer cuáles son los «puntos distantes» es a través de la distancia:

$$D_i^2 = \sum_{j=1}^{k} \left(\frac{b_j(x_{ij} - \overline{x}_j)}{\sqrt{MCE}} \right)^2, \qquad i = 1, 2, \ldots, n.$$

El procedimiento consiste en ajustar el modelo, calcular los D_i^2, $i = 1, 2, \ldots, n$, y después ordenarlos en forma ascendente de acuerdo con el D_i^2. Los puntos con alto D_i^2 son inusuales.

Existen otros tipos de distancias que permiten la detección de valores atípicos y puntos influyentes, cada una con distintas propiedades, pero todas siguen el mismo principio para la identificar tales observaciones.

Tratamiento

Una vez localizado un punto inusual, se estudiará su influencia eliminándolo del modelo, repitiendo el procedimiento para todos estos puntos. Aquí es necesario realizar un análisis de la estabilidad de los coeficientes de regresión mediante pruebas estadísticas o sus intervalos de confianza.

También, se han desarrollado *métodos de regresión robusta* que atenúan la influencia de las observaciones inusuales en el modelo o que toman como medida estadística sobre la cual se basa la regresión la mediana en lugar de tomar la media, como lo hemos hecho durante este tratado.

La transformación de la variable dependiente a una forma que sea más cercana a la normal es un recurso muy empleado. Estas transformaciones suelen ser sugeridas por los gráficos de los residuos. También, se puede utilizar el método de Box-Cox de transformación potencial.

En muchos casos, la desviación de la normalidad se debe a la presencia de valores atípicos, en cuyo caso es conveniente examinar la influencia de los mismos.

Recientemente se han desarrollado modelos que consideran que los errores están distribuidos según una ley t, de un número de grados de libertad desconocido, como una generalización de la hipótesis de normalidad.

11.8.3. Presencia de puntos inusuales

Generalmente, se espera que los datos correspondientes a las observaciones se encuentren distribuidos en una región más o menos cercana, pero puede suceder que una o varias observaciones estén alejadas del resto de los datos. Estas observaciones pueden influir mucho en el modelo final.

Este aspecto es muy importante: podemos disponer de 100 observaciones y, sin embargo, construir con ellos un modelo cuyas propiedades son debidas únicamente a dos puntos. Conocer si este tipo de puntos influye perjudicialmente en el modelo permite mejorarlo.

Identificación

La primera forma para determinar si un valor es atípico es mediante los residuos estudentizados. Se compara con los valores de t_i con los valores críticos de una ley t con $n-k-2$ grados de libertad.

Otra forma de conocer cuáles son los «puntos distantes» es a través de la distancia:

$$D_i^2 = \sum_{j=1}^{k} \left(\frac{b_j(x_{ij} - \overline{x}_j)}{\sqrt{MCE}} \right)^2, \qquad i = 1, 2, \ldots, n.$$

El procedimiento consiste en ajustar el modelo, calcular los D_i^2, $i = 1, 2, \ldots, n$, y después ordenarlos en forma ascendente de acuerdo con el D_i^2. Los puntos con alto D_i^2 son inusuales.

Existen otros tipos de distancias que permiten la detección de valores atípicos y puntos influyentes, cada una con distintas propiedades, pero todas siguen el mismo principio para la identificar tales observaciones.

Tratamiento

Una vez localizado un punto inusual, se estudiará su influencia eliminándolo del modelo, repitiendo el procedimiento para todos estos puntos. Aquí es necesario realizar un análisis de la estabilidad de los coeficientes de regresión mediante pruebas estadísticas o sus intervalos de confianza.

También, se han desarrollado *métodos de regresión robusta* que atenúan la influencia de las observaciones inusuales en el modelo o que toman como medida estadística sobre la cual se basa la regresión a la mediana en lugar de tomar la media, como lo hemos hecho durante este tratado.

11.8.4. Varianza heterogénea

La heterocedasticidad en los errores implica que la hipótesis

$$\mathrm{Var}(e_i) = \sigma^2 = \text{constante}$$

no es aplicable.

Las consecuencias de la heterocedasticidad en el modelo lineal son las siguientes: los estimadores serán insesgados, pero dejan de ser eficientes. Las fórmulas para las varianzas no son correctas y las pruebas de hipótesis dejan de ser aplicables.

Identificación

La heterocedasticidad se identifica mediante la graficación de los residuos. El gráfico $e_i = f(\widehat{y}_i)$ es útil para detectarla y los gráficos $e_i = f(x_{ki})$ para identificar si la variabilidad es causada por alguna de las variables independientes en particular.

También es posible realizar una prueba de hipótesis del tipo

$$H_0\colon e_i \sim \mathcal{N}\left(0, \sigma^2\right),$$
$$H_1\colon e_i \sim \mathcal{N}\left(0, \sigma_i^2\right),$$

que es una prueba de ajuste a una ley. Para el efecto se puede realizar el contraste de razón de verosimilitudes, que se basa en el estadístico

$$2\log\lambda = n\log\widehat{\sigma}^2 - \sum n_i \log\widehat{\sigma}_i^2,$$

que se distribuye según una ley χ^2 con $k-1$ grados de libertad.

Tratamiento

1. Si la heterocedasticidad está asociada con la variable respuesta y crece con el incremento de los valores de y, una posible forma de tratarla es realizar la regresión de $\log y$ en lugar de y. Este caso suele aparecer cuando hay una formulación errónea del modelo; por ejemplo cuando el modelo real es multiplicativo y se ajusta mediante uno lineal.

2. Si la heterocedasticidad está asociada con una variable independiente particular y la desviación estándar está ligada linealmente con el crecimiento de x_k, un procedimiento de evitarla es estimar el modelo
$$\frac{y}{x_k} = \frac{\beta_0}{x_k} + \beta_1\frac{x_1}{x_k} + \cdots + \beta_k + \varepsilon^*,$$
donde, al dividir todos los términos por x_k, los residuos $\varepsilon^* = \dfrac{\varepsilon}{x_k}$ tendrán varianza constante. Este procedimiento es equivalente a utilizar mínimos cuadrados generalizados.

11.8.5. Multicolinealidad

En los problemas de regresión múltiple, algunas veces, dos o más variables independientes contribuyen con información redundante, porque se encuentran bastante correlacionadas entre sí. En los casos en que tal correlación sea alta, se dice que existe *multicolinealidad*. Por ejemplo, se desea formar un modelo para predecir el precio del metro cuadrado de tierra en un sector de la ciudad (y), como función del índice de inflación (x_1) y del precio del dólar en el mercado libre (x_2). Aunque las dos variables aportan información, es conocido que las dos variables independientes están fuertemente correlacionadas; por lo tanto, la información por ellas aportada es redundante o se traslapa.

En este caso, la matriz $\mathbf{X}^t\mathbf{X}$ es casi singular, originando que sea difícil encontrar su inversa.

Cuando hay multicolinealidad se tiene dos efectos:

1. Los estimadores b_i tienen varianzas muy altas.
2. Las estimaciones b_i son muy dependientes entre sí.

Por ejemplo, en un modelo con dos variables independientes,

$$\text{Var}(b_i) = \frac{\sigma^2}{s_i^2\,(1 - r^2)\,n},$$

donde r es el coeficiente de correlación entre las dos variables. Así, si aumenta (en valor absoluto) la correlación entre las variables explicativas, aumentarán las varianzas de las estimaciones y su dependencia.

Identificación

La identificación de variables colineales se efectúa examinando:

1. La matriz de correlación entre las variables explicativas, $\mathbf{R}$, y su matriz inversa, $\mathbf{R}^{-1}$.
2. Las raíces y vectores propios de las matrices $\mathbf{X}^t\mathbf{X}$, o de su matriz de correlación.

Si una variable es combinación lineal de las restantes variables, se debe analizar la matriz $\mathbf{R}^{-1}$. Así, se define el «factor de inflación de la varianza» como el $i-$ésimo término de la diagonal de la matriz $\mathbf{R}^{-1}$:

$$FIV_i = \text{diag}_i\left(\mathbf{R}^{-1}\right).$$

Por tanto, elementos diagonales grandes (mayores a 10) en la matriz $\mathbf{R}^{-1}$ indican alta colinealidad.

También, se puede calcular el índice de condicionamiento (IC) de $\mathbf{X}^t\mathbf{X}$ o $\mathbf{R}$:

$$\sqrt{\frac{\text{máx}\,\{\lambda_i\}}{\text{mín}\,\{\lambda_i\}}} \geq 1,$$

donde λ_i son los valores propios de las mencionadas matrices.

En la práctica se admite que existe alta multicolinealidad cuando el IC es mayor que 30; una colinealidad moderada si el IC está entre 10 y 30; y, cuando el IC es menor que 10 se considera que la matriz está bien definida.

Tratamiento

Las tres soluciones a la multicolinealidad son:

1. Eliminar regresores, reduciendo el número de parámetros a estimar.
2. Transformar las variables mediante componentes principales y elimimar los menos importantes.
3. Incluir información externa a los datos, mediante un enfoque bayesiano.
4. Utilizar regresión a través de componentes principales, que es similar al método de eliminar regresores, con la diferencia de que se eliminan combinaciones lineales de éstos, manteniendo aquellos cuya contribución en términos de información es significativa.

11.8.6. Autocorrelación

Una de las hipótesis iniciales para desarollar el modelo de regresión es que los errores, ε_i, son variables aleatorias no correlacionadas. Si esta hipótesis es violada, se dice que existe *autocorrelación*.

Los efectos de esta dependencia son los siguientes:

1. Los estimadores de los parámetros son insesgados, pero no son eficientes.
2. Los contrastes sobre los parámetros no son válidos y están sesgados hacia la detección de relaciones inexistentes.
3. Las predicciones son ineficientes.

En el caso de un modelo lineal simple, la varianza del estimador de la pendiente resulta ser

$$\text{Var}(b) = \frac{\sigma^2}{\sum x_i^2}\left(1 + 2\rho\frac{\sum x_i x_{i+1}}{\sum x_i^2}\right),$$

donde ρ es el coeficiente de correlación entre las observaciones.

Si $\rho > 0$, la varianza puede resultar sustancialmente mayor que $\frac{\sigma^2}{\sum x_i^2}$, de manera que el estimador es ineficiente.

Identificación

Para la detección de la autocorrelación se emplea el estadístico de Durbin-Watson:

$$r_h = \frac{\sum e_i e_{i-h}}{\sum e_i^2}.$$

Para la realización de las pruebas estadísticas sobre la autocorrelación existen tablas, pero podemos dar la siguiente regla empírica para detectar la autocorrelación de orden 1.

Construimos el estadístico

$$d = 2(1 - r_1)$$

y si su valor es próximo a 2, no existirá autocorrelación; si d tiene un valor entre 0 y 2, habrá correlación positiva; y, si el valor de d está entre 2 y 4, la correlación será negativa. En los programas informáticos se suele calcular el valor de este estadístico y el nivel de significación de la prueba.

Alternativamente, se puede aplicar el estadístico de Box-Ljung para detectar la autocorrelación.

Tratamiento

Para explicar la evolución de variables que tienen un comportamiento en el que aparece autocorrelación, es conveniente utilizar métodos del análisis de series de tiempo, que permiten abordar de manera más global el problema de construcción de modelos para estas variables. También, se pueden utilizar métodos especiales de regresión, como los mínimos cuadrados generalizados o los modelos lineales generalizados.

Por la dificultad que entraña la realización manual de los cálculos, especialmente en la determinación de los potenciales problemas que el modelo pudiera presentar, ésto se hace mediante el empleo de programas estadísticos especializados, que facilitan su cálculo, correspondiendo al usuario la interpretación correcta de los resultados.

El lector deberá notar que los temas tratados aquí solo cubren la parte central del análisis de regresión. Existen textos especializados que lo tratan de manera detallada y en extenso. (Véase Rawlings y otros, 2001.)

11.9. Ejercicios

Modelos de regresión múltiple

1. Suponga que un investigador está usando el modelo estadístico $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, donde $n = 13$ y $k = 2$. Resultó el siguiente sistema de ecuaciones, que permiten estimar el vector $\boldsymbol{\beta}$:

$$\begin{aligned} 13b_0 + 2b_1 - 2b_2 &= 4 \\ 2b_0 + 2b_1 - b_2 &= 2 \\ -2b_0 - b_1 + 4b_2 &= -1 \end{aligned}$$

   a) Resuelva el sistema de ecuaciones y encuentre el estimador $\mathbf{b}$ de coeficientes. Escriba explícitamente la ecuación de regresión;
   b) Escriba la formulación matricial del problema;
   c) Si $\mathbf{Y}^t\mathbf{Y} = 6$, encuentre s^2;
   d) Calcule los coeficientes de determinación múltiple R^2 y ajustado R_a^2 e interprételos.

2. Una empresa de transporte ha tomado una muestra de los pesos de seis embarques, la distancia transportada y el gasto que en ellos se ha incurrido:

Peso (Tm)	4.0	3.0	1.6	1.2	3.4	4.8
Distancia (miles de km)	2.5	2.2	1.0	2.0	0.8	1.6
Gastos (dólares)	320	224	138	180	246	372

   a) Estime los coeficientes β_0, β_1 y β_2 del modelo de regresión lineal múltiple;
   b) Use el modelo para predecir el gasto cuando el embarque pesa 2.4 Tm y se lo transporta a 1200 km;
   c) Calcule e interprete el coeficiente de determinación múltiple y el coeficiente de determinación ajustado. Comente los resultados.

3. Un economista está interesado en la relación que existe entre la demanda de viviendas, su precio y el ingreso medio anual de los hogares. Si denominamos por y la demanda de vivienda, medida en unidades adecuadas; x_1 al precio promedio de las viviendas; y, x_2 el ingreso familiar promedio. Los valores de estas variables se recogieron para 6 periodos y se muestran en la siguiente tabla:

Periodo	y	x_1	x_2
1	8	12	6.8
2	9	13	7.2
3	12	13	7.4
4	9	14	7.1
5	12	14	7.0
6	15	15	7.4

Si el economista supone que el modelo apropiado es uno de regresión lineal múltiple:

a) estime los coeficientes de regresión del modelo;

b) calcule el coeficiente de determinación;

c) realice la tabla de análisis de la varianza e interprete su resultado;

d) efectúe las pruebas para determinar la significación de los parámetros individuales.

4. En la tabla se dan los datos sobre las ventas semanales promedio (en cientos de dólares) de cada uno de los restaurantes que tiene una cadena de comida rápida, su capacidad de asientos y el número de clientes (en cientos) que ingresan a cada local.

Número de asientos	Número de clientes	Ventas semanales
120	19	13.8
200	8	14.2
150	12	14.0
180	15	18.2
240	16	23.5

a) Asumiendo que el modelo de regresión es lineal múltiple, estime los coeficientes β_0, β_1 y β_2 y sus intervalos de confianza correspondientes (use 95 %);

b) Calcule R^2 y R_a^2. Interprete los valores obtenidos;

c) Construya la tabla ANOVA y realice la prueba de adecuación del modelo;

d) Use el modelo para realizar la estimación y la predicción de la venta semanal de un restaurante que será instalado con una capacidad de 150 asientos y se espera que ingresen 1400 clientes.

5. De una encuesta de presupuestos familiares se han obtenido los siguientes datos mensuales:

Gasto en energía eléctrica ($)	30	40	50	70	70	130
Ingreso familiar ($)	300	800	1500	2400	1200	3000
Tamaño familiar	2	4	3	6	4	8

a) Construya e interprete un modelo para explicar el gasto en energía eléctrica en función del ingreso familiar y el tamaño de la familia;

b) Calcule los coeficientes de determinación (múltiple y ajustado) y la varianza residual;

c) Construya un intervalo de confianza con nivel 95 % para el efecto de la variable ingreso familiar;

d) Indique qué coeficientes son significativos;

e) Si el tamaño de la familia permanece constante, calcule la variación del gasto esperado para un incremento de renta de 350 dólares.

f) Efectúe la prueba sobre la significación de la regresión.

6. Sean las variables: L = latitud en grados, A = altura en metros sobre el nivel del mar y T = temperatura media anual.

a) Explique T en función de L y A con base en la información obtenida en varias ciudades:

L	33.4	33.2	31.3	29.5	26.8	26.5
A	122	145	195	124	107	130
T	13.9	14.9	16.4	17.2	18.0	18.0

b) Prevea la temperatura media para una ciudad cuya latitud es 30.5 y la altitud es 150 m;

c) Calcule los coeficientes de determinación múltiple y ajustado;

d) Realice la tabla de análisis de la varianza e interprétela;

e) Contraste la nulidad de cada uno de los parámetros de la regresión e indique si es posible encontrar un modelo mejor que el planteado.

7. Se estimó un modelo de regresión múltiple a partir de 25 observaciones y se obtuvo:

Coeficiente	**Estimación**	**Error estándar**
β_0	10.6	2.1
β_1	28.4	11.2
β_2	4.0	1.5
β_3	12.7	14.1
β_4	0.84	0.76

a) En el modelo propuesto: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$, ¿qué regresores son estadísticamente significativos? (Use un nivel del 5 %);

b) Usando los resultados de a), prediga el valor de la respuesta para $x_p = (1, 2, 1, -1)^t$ mediante un intervalo.

8. En un estudio sobre la relación entre tres variables se obtuvieron 11 mediciones con los siguientes resultados

$$\sum_{i=1}^{11} x_{1i} = 66, \quad \sum_{i=1}^{11} x_{1i}^2 = 506, \quad \sum_{i=1}^{11} x_{1i} y_i = 85,$$
$$\sum_{i=1}^{11} x_{2i} = -22, \quad \sum_{i=1}^{11} x_{2i}^2 = 484, \quad \sum_{i=1}^{11} x_{2i} y_i = 142,$$
$$\sum_{i=1}^{11} y_i = 33, \quad \sum_{i=1}^{11} y_i^2 = 289, \quad \sum_{i=1}^{11} x_{1i} x_{2i} = -346.$$

a) Halle la ecuación de regresión lineal que ajusta los datos;

b) Realice la tabla de análisis de la varianza e interprétela;

c) Encuentre los intervalos de confianza de los parámetros del modelo e interprételos. (Use $1 - \alpha = 0.9$);

d) Halle el coeficiente de determinación entre las variables.

Regresión polinomial

9. En el laboratorio de una empresa automotriz se midió la distancia que tarda un automóvil en frenar, de acuerdo a la velocidad que lleva el momento que aparece una señal. Los datos se muestran en la siguiente tabla:

Velocidad (km/h)	30	40	60	80	90	100	110	120
Distancia (m)	10	20	30	50	70	90	150	250

a) Ajuste una ecuación polinomial de segundo grado a los datos;

b) Determine la calidad del ajuste;

c) Realice una prueba global de la bondad de ajuste mediante la tabla ANOVA.

10. En una entidad bancaria las tasas efectivas de interés varían de acuerdo al monto del préstamo que se concede, ya que se incluyen gastos administrativos e impuestos que cobra el estado. Para encontrar un modelo, se recogieron los siguientes datos, de varios préstamos concedidos:

Monto (miles de $)	0.5	1.0	1.5	2.5	4.0	6.0	9.0	12.0
Tasa (%)	23	23	21	17	15	16	19	21

a) Grafique los datos y postule un modelo de regresión para los datos;

b) Ajuste los datos mediante el modelo de regresión de a);

c) Realice una prueba global de adecuación del modelo;

d) Pruebe si los términos individuales del modelo pueden eliminarse.

11. Los datos que se presentan a continuación corresponden al tiempo de secado de un barniz y la cantidad de cierto aditivo químico añadido:

Cantidad de aditivo (g)	1	2	3	4	5	6	7	8
Horas de secado	8.5	8.0	6.0	5.0	6.0	5.5	6.5	7.0

a) Use el método de ajuste a un modelo polinomial de segundo orden para modelizar los datos;

b) Realice una predicción para el tiempo de secado, cuando se han añadido 6.5 gramos de aditivo químico;

c) Calcule los coeficientes de determinación R^2 y R_a^2 e interprételos.

12. En un estudio de la contaminación por gases despedidos por los vehículos en las calles de Quito, se midió la concentración de plomo (en ppm) en el aire y se registró la temperatura ambiente (en °C) el momento de la medición.

Temperatura	8.0	10.1	12.6	14.3	15.6	16.2	17.1	18.7
Concentración	73	68	74	88	97	107	115	118

a) Encuentre una ecuación de segundo orden que ajuste las observaciones;

b) Realice un análisis de varianza de la regresión y pruebe la significación del ajuste;

c) ¿Es posible simplificar el modelo eliminando una de las componentes? ¿Por qué?

Regresión con variables ficticias

13. Se cree que en ciertas empresas existe discriminación sexual en el trabajo y que se paga más a los hombres que a las mujeres cuando realizan el mismo trabajo. Los datos de la tabla corresponden a los años de experiencia, el sexo y los sueldos anuales (en miles de dólares), de empleados que tienen las mismas funciones en 9 empresas de consultoría.

Experiencia (años)	Sexo	Sueldo anual (miles)
3	mujer	12.9
4	mujer	13.0
5	mujer	13.8
6	mujer	14.4
8	mujer	15.7
4	hombre	13.4
5	hombre	14.1
7	hombre	17.4
8	hombre	21.1

a) Ajuste un modelo de regresión a los datos;

b) Pruebe la significación de la regresión;

c) Pruebe las hipótesis H_0: $\beta_i = 0$ para los coeficientes del modelo;

d) Utilizando el resultado anterior, ¿se puede concluir que existe discriminación? ¿Por qué?

14. En una empresa que se dedica a la transportación de turistas, se sospecha que el consumo de combustible de los autos depende de la antigüedad de éstos y del sexo de quien los conduce. Para probar esta sospecha se recogió la siguiente información:

Consumo (gal/100 km)	Sexo del conductor	Antigüedad del auto
4.1	H	3
5.0	M	4
4.6	M	3
4.0	H	2
5.2	M	1
4.7	M	4
3.5	H	4
4.7	H	1

a) Con la ayuda de variables indicadoras ajuste un modelo de regresión a los datos;

b) Elabore una tabla de análisis de la varianza y calcule el coeficiente de determinación múltiple;

c) Realice pruebas sobre los parámetros de la regresión y determine cuál de ellos debe ser excluido del modelo;

d) Interprete los resultados obtenidos.

15. Los siguientes datos corresponden al tiempo de sobrevivencia, en horas, de unas ratas luego que se les suministra cierta dosis de veneno. En el experimento se consideró 2 tipos de veneno (A y B) y la edad de la rata.

Edad (meses)	1	2	3	4	1	2	3	4	2	3
Veneno	A	A	A	A	B	B	B	B	A	B
Sobrevivencia	4.5	8.2	6.3	7.1	3.6	4.9	4.4	5.6	6.7	5.0

a) Realice un ajuste de los datos con la ayuda de variables indicadoras;

b) Pruebe la significación de la regresión y calcule el coeficiente de determinación múltiple;

c) Calcule los estadísticos t para los coeficientes del modelo y examine si alguna de las variables puede excluirse;

d) De acuerdo a lo obtenido en c), encuentre el modelo definitivo.

Capítulo 12

Introducción a las Series de Tiempo

El pronóstico es el arte de establecer lo que hubiera sucedido
si lo que sucedió no hubiera sucedido.
David Myddleton

Para el análisis de las series de tiempo existen varios enfoques, ninguno de los cuales se puede decir que es mejor que los restantes. Nosotros nos referiremos al más simple, que es el de identificar los principales factores que parecen influir en los valores individuales de la misma y examinaremos los métodos de predicción que se agrupan bajo el nombre genérico de métodos de suavizamiento o de atenuación.

12.1. Introducción

Para el análisis de una variedad de fenómenos físicos, económicos o financieros disponemos de una cierta cantidad de observaciones de una misma variable aleatoria numérica, tomadas en momentos equidistantes; entonces, los datos analizados tienen propiedades interesantes algunas de las cuales las vamos a presentar.

Primero, definamos lo que se entiende por serie de tiempo (que también se denomina como *serie histórica* o *serie cronológica*).

Definición (de serie de tiempo) Una serie de tiempo es un conjunto de datos que están ordenados en el tiempo y que han sido tomados a intervalos equidistantes en el tiempo.

Como ejemplos de series cronológicas podemos mencionar los siguientes:

1. La temperatura diaria promedio tomada en un lugar específico, durante un año.
2. El índice mensual de precios al consumidor (IPC).
3. Los precios de venta de ciertos artículos, de temporada.
4. El caudal diario promedio de un río, en un sitio determinado.
5. El número de empleados en ramas relacionadas con el turismo.

Si las observaciones de un fenómeno se han realizado a través del tiempo, es posible que presenten esquemas que se repiten periódicamente y que las observaciones contiguas sean, probablemente, dependientes.

A las series cronológicas se las representa mediante diagramas de dispersión (gráficos $X-Y$) donde las observaciones se representan en el eje de las ordenadas y la variable temporal en el eje de las abscisas. La variable tiempo (t) es la variable independiente.

Nosotros llevaremos la siguiente notación:

1. La variable tiempo (t), que tomará valores en los enteros positivos: $t = 0, 1, 2, \ldots, n$.
2. La variable de las observaciones (Y), que la supondremos dependiente del tiempo: Y_0, Y_1, ..., Y_n.

Para realizar el análisis, en primera instancia, se aplicarán los métodos tradicionales de la inferencia estadística para modelizar los datos y conocer la bondad del ajuste realizado.

12.2. Componentes de una serie de tiempo

Desde el punto de vista tradicional, los componentes de una serie de tiempo son: la tendencia secular, y los componentes estacional, cíclico e irregular.

12.2.1. La tendencia secular

De los componentes que afectan a los valores individuales en una serie cronológica, el más importante es, generalmente, la *tendencia secular* (o llamada simplemente tendencia), que se define como el movimiento característico de crecimiento (o de decrecimiento) a largo plazo de la serie. Por eso, la tendencia solo se puede determinar si se dispone de datos de la serie en un periodo largo de años.

La tendencia es la dirección subyacente (hacia arriba o hacia abajo) en la serie de datos, a largo plazo.

A la tendencia se la identifica mediante el símbolo T.

Las fuerzas básicas que producen o afectan la tendencia son: los cambios en la población, la inflación, el cambio tecnológico, el incremento de la productividad y los ciclos de vida de los productos.

Para la modelización de este componente se utilizan modelos lineales, polinomiales y otros. El método más simple, y más ampliamente usado para describir la tendencia es la regresión lineal simple y las transformaciones que pueden hacerse hacia este modelo. Así, la tendencia puede ser lineal –creciendo a una cantidad absoluta constante a lo largo del tiempo– o puede ser exponencial, creciendo a una tasa constante a lo largo del tiempo. Alternativamente, la tendencia puede ajustarse a un polinomio o otro modelo más complejo.

Por su naturaleza el análisis de la tendencia tiene implicaciones en la planificación administrativa a largo plazo.

12.2.2. La variación estacional

El componente estacional de la serie es un patrón de cambio que se repite regularmente en el tiempo.

Este movimiento debe completarse dentro del periodo de un año y repetirse, de manera semejante año tras año, a fin de considerarse un cambio estacional. Así, para identificar el componente estacional en una serie histórica, es necesario recopilar los datos para más de un periodo de un año.

Por ejemplo, si consideramos los registros de demanda de habitaciones en los hoteles de los sitios turísticos más visitados, durante el año. En los meses de vacaciones de los regímenes escolares de la Costa y de la Sierra se produce mayor demanda que en el resto de meses del año. Así, los datos presentarán variaciones estacionales, con una marcada tendencia a aumentar, durante los periodos señalados.

Mientras que la tendencia se utiliza para la planificación a largo plazo, el análisis del componente estacional de una serie histórica tiene implicaciones a corto plazo, más inmediatas.

Las fluctuaciones estacionales se presentan típicamente en los datos clasificados por meses o trimestres; lo que conduce a que se deba calcular un valor estacional por separado para cada mes (o trimestre) del año, por lo general en la forma de un número índice.

En la Figura 12.1 se gráfica una serie con comportamiento estacional. La variación estacional se representa mediante E.

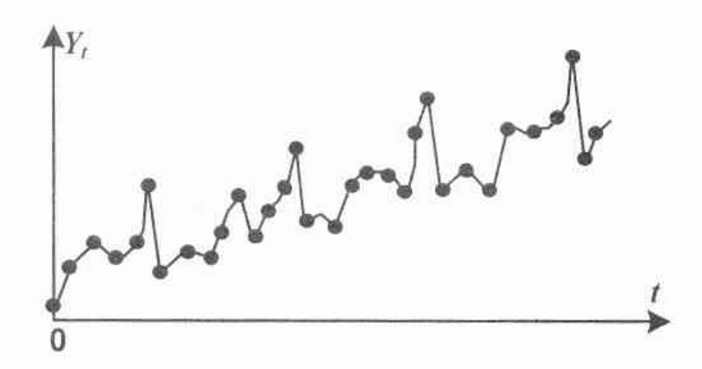


Figura 12.1: Serie de tiempo con fluctuaciones estacionales.

12.2.3. La variación cíclica

El componente cíclico es la fluctuación en forma de ondas o ciclos, de más de un año de duración, producido por cambios en las condiciones económicas.

Los componentes cíclicos se asemejan a los estacionales en que también son movimientos ondulatorios repetitivos, pero difieren en que los movimientos son de duración más prolongada y son menos predecibles en duración y en amplitud.

El análisis del componente cíclico es de importancia en la planificación a largo y mediano plazos ya que permite preveer los periodos en los que estas variaciones afectarán las actividades.

Un ejemplo de variación cíclica se presenta en el precio del petróleo, que en las últimas décadas ha tenido periodos largos de crecimiento sostenido, seguidos de periodos de fuerte caída.

La variación cíclica se la identifica por C.

12.2.4. La fluctuación irregular

El componente irregular corresponde a fluctuaciones causadas por sucesos impredecibles o no periódicos.

El componente irregular puede deberse a fenómenos como un clima poco usual, huelgas, guerras, elecciones y cambios en las leyes, o por los errores que se cometen al realizar las mediciones.

El componente irregular de la serie histórica se identifica con el símbolo I.

El principal uso de las series de tiempo es la realización de pronósticos. En este capítulo nos concentraremos en un conjunto de técnicas de pronósticos conocidas como *métodos de atenuación* de las series, que son fáciles de implementar y no requieren de técnicas matemáticas sofisticadas.

12.3. Atenuación de series de tiempo

Las técnicas de atenuación o *suavizamiento* se emplean para reducir las irregularidades; es decir, las fluctuaciones aleatorias, en una serie de tiempo, proveyendo una visión más clara del comportamiento subyacente en la serie de datos.

En algunas series cronológicas la variación estacional es tan pronunciada que no permite apreciar la tendencia o los ciclos, que son muy importantes para entender el proceso observado. El suavizamiento puede remover la estacionalidad y hace que las fluctuaciones a largo plazo aparezcan más claras.

Además, con frecuencia, el analista quiere actualizar los pronósticos diaria, semanal o mensualmente de manera rápida, barata y sencilla, eso se puede lograr mediante los métodos de suavizamiento de los datos.

Las técnicas más comunes son el suavizamiento por *promedios móviles* y el *suavizamiento exponencial.* Como el tipo de estacionalidad varía de serie a serie, así debe variar el tipo de suavizamiento empleado.

Estas técnicas se basan en promedios ponderados de mediciones anteriores. La explicación consiste en que los valores pasados contienen información de lo que ocurrirá en el futuro. Debido a que los valores anteriores incluyen fluctuaciones aleatorias, así como información relativa al patrón subyacente de la variable en estudio, se hace un intento de suavizar estos valores.

Estos métodos son apropiados cuando:

1. Hay que realizar pronósticos de muchas series.
2. Solamente se requieren pronósticos a corto plazo (hasta unos cuantos meses).
3. Es aceptable una precisión razonable, mas no exacta.
4. Desde el punto de vista costo/beneficio, no se justifican procedimientos más complicados.

Teniendo en cuenta lo anterior, se pueden enumerar brevemente las ventajas y desventajas de los métodos de pronóstico por suavizamiento:

Ventajas

1. Simplicidad, facilidad de comprensión.

2. Precisión aceptable (pero de ninguna manera exactitud total) en una amplia variedad de aplicaciones.

3. Fácil implementación informática.

Desventajas

1. Por lo general, requieren de un sistema hombre/máquina, con vigilancia manual y posibilidades de invalidación.

2. No pronostican los puntos cruciales, cuando hay cambios bruscos en el patrón de los datos.

3. Se aplican solo en los pronósticos a corto plazo.

4. Pueden reaccionar excesivamente al azar, causando oscilaciones destructivas.

5. Pueden presentar problemas técnicos en la selección del modelo correcto y en la selección de las constantes de suavizamiento.

12.3.1. Suavizamiento por promedios móviles

Los datos históricos se pueden atenuar en muchas formas, pero siempre con el objetivo de usar los datos disponibles para desarrollar un modelo de pronóstico para periodos futuros.

El método más simple e intuitivo es usar el promedio simple, consistente en encontrar la media aritmética de todas las observaciones y usarlo para pronosticar el siguiente periodo; es decir,

$$\widehat{Y}_{t+1} = \frac{\sum_{t=1}^{n} Y_t}{n}.$$

Este método, en general, tiene una validez muy limitada ya que es aplicable cuando los datos no presentan tendencia, estacionalidad u otros patrones sistemáticos.

Ejemplo. En el Cuadro 12.1 se presentan las observaciones que corresponden a las ventas trimestrales de una empresa en el periodo 2003 a 2006.

Pronosticar las ventas para el año 2007.

Solución: El promedio de los 16 datos es igual a 556.6. Es decir, $\widehat{Y}_{17} = 556.6$.

El valor del pronóstico puede parecer bajo ya que en los primeros trimestres de los últimos años las ventas fueron mayores que el valor pronosticado. Además, si deseamos un pronóstico para cada uno de los trimestres del año 2007, no podemos hacerlo. Vemos que el método propuesto es insuficiente para nuestros propósitos. ◄

En muchos análisis, los datos que presentan mayor interés son los más recientes ya que ellos darán mayor información del actual estado del proceso que aquellos datos que fueron tomados al inicio del mismo. Para realizar ésto se puede calcular el promedio para las observaciones más recientes. Al estar disponible una nueva observación, se puede calcular una nueva media eliminando el valor más antiguo e incluyendo el más reciente.

Para describir este enfoque se emplea el término promedio (o media) móvil. Entonces, se usa este promedio móvil para pronosticar el siguiente periodo.

Año	Trimestre	t	Ventas
2003	1	1	598
	2	2	390
	3	3	267
	4	4	573
2004	1	5	588
	2	6	425
	3	7	371
	4	8	609
2005	1	9	777
	2	10	532
	3	11	433
	4	12	689
2006	1	13	855
	2	14	618
	3	15	460
	4	16	720

Cuadro 12.1: Datos correspondientes a las ventas trimestrales de una empresa.

La expresión matemática de la media móvil es

$$M_t = \widehat{Y}_{t+1} = \frac{Y_t + Y_{t-1} + Y_{t-2} + \cdots + Y_{t-n+1}}{n},$$

donde

$M_t =$ promedio móvil en el periodo t;

$\widehat{Y}_{t+1} =$ valor del pronóstico para el siguiente periodo;

$Y_t =$ valor real en el periodo t; y,

$n =$ número de términos en el promedio móvil.

Así, el promedio móvil para el periodo t es la media aritmética de las n observaciones más recientes.

Observemos que el número de periodos que intervienen en el cálculo de una media móvil particular no cambia al correr del tiempo. Por ésto, es importante escoger el número de periodos n, que se denomina su orden. Para datos trimestrales es frecuente que se utilice un promedio móvil de orden 4 y para datos mensuales un promedio móvil de orden 12; es decir, se emplea el mismo orden de la estacionalidad de la serie.

Ejemplo. (Continuación) Realizar el pronóstico, mediante medias móviles, de los datos del Cuadro 12.1.

Solución: Como los datos están dados en forma trimestral tomaremos una media móvil de orden 4. En el Cuadro 12.2 se encuentran los pronósticos para la serie completa.

Examinemos los cálculos del pronóstico del periodo 17:

$$\begin{aligned} M_{16} &= \widehat{Y}_{17} = \frac{Y_{16} + Y_{15} + Y_{14} + Y_{13}}{4} \\ &= \frac{720 + 460 + 618 + 855}{4} = 663.3. \end{aligned}$$

Entonces, para el primer trimestre del año 2006 se espera tener un nivel de ventas igual a $\widehat{Y}_{17} =$ 663.3. Si comparamos con las observaciones de los últimos años, el valor del pronóstico es mucho más razonable que el anteriormente obtenido, pero persiste el problema de la realización de pronósticos a mayor plazo.

Año	Trimestre	t	Ventas Y_t	Pronostico de Promedio movil, $\widehat{Y}_t$
2002	1	1	598	
	2	2	390	
	3	3	267	
	4	4	573	
2003	1	5	588	457.1
	2	6	425	454.5
	3	7	371	463.2
	4	8	609	489.3
2004	1	9	777	498.2
	2	10	532	545.5
	3	11	433	572.4
	4	12	689	587.7
2005	1	13	855	607.8
	2	14	618	627.3
	3	15	460	648.8
	4	16	720	655.6
2006	1	17		663.3

Cuadro 12.2: Pronóstico mediante promedio móvil.

12.3.2. Suavizamiento por la media móvil doble

Es una variante del método expuesto, que tiene mayor efectividad cuando se dispone de una serie de tiempo cuya tendencia es lineal. El método consiste en, primero, calcular un conjunto de promedios móviles; y luego, se calcula un segundo conjunto como promedio móvil del primero. La técnica de atenuación por promedio móvil doble se realiza de la forma siguiente:

1. Se calcula el promedio móvil (simple) de los datos:
$$M_t = \frac{Y_t + Y_{t-1} + Y_{t-2} + \cdots + Y_{t-n+1}}{n}. \tag{12.1}$$

2. Dado que $M_t = \widehat{Y}_{t+1}$, se usa esta nueva serie para calcular un segundo conjunto de promedios móviles:
$$M'_t = \frac{M_t + M_{t-1} + M_{t-2} + \cdots + M_{t-n+1}}{n}. \tag{12.2}$$

3. Se calcula la diferencia entre los dos promedios móviles:
$$a_t = 2M_t - M'_t \tag{12.3}$$
y un factor de ajuste adicional, similar a la medición de una pendiente que cambia a través de la serie:
$$b_t = \frac{2}{n-1}\left(M_t - M'_t\right). \tag{12.4}$$

4. Se forma la ecuación que servirá para realizar el pronóstico en p periodos en el futuro
$$\widehat{Y}_{t+p} = a_t + b_t p, \tag{12.5}$$

Ejemplo. (Continuación) Realizar el pronóstico, con la técnica de la promedio móvil doble, de los datos del Cuadro 12.1.

Solución: Los resultados se presentan en el Cuadro 12.3.

A continuación se exponen los cálculos para comprender el pronóstico del trimestre 17.

Periodo t	Ventas Y_t	Promedio móvil	Promedio móvil doble	Valor de a	Valor de b	Pronóstico $a + bp$ $(p = 1)$
1	598					
2	390					
3	267					
4	573	457.1				
5	588	454.5				
6	425	463.2				
7	371	489.3	466.0	512.5	15.5	
8	609	498.2	476.3	520.1	14.6	528.0
9	777	545.5	499.1	592.0	31.0	534.7
10	532	572.4	526.3	618.4	30.7	623.0
11	433	587.7	550.9	624.4	24.5	649.1
12	689	607.8	578.3	637.3	19.6	648.9
13	855	627.3	598.8	655.8	19.0	656.9
14	618	648.8	617.9	679.8	20.6	674.8
15	460	655.6	634.9	676.3	13.8	700.4
16	720	663.3	648.8	677.8	9.7	690.1
17						687.5

Cuadro 12.3: Pronóstico por promedio móvil doble.

1. Se calculan los promedios móviles de cuatro trimestres mediante la ecuación 12.1:

$$\begin{aligned} M_{16} &= \frac{Y_{16} + Y_{15} + Y_{14} + Y_{13}}{4} \\ &= \frac{855 + 618 + 460 + 720}{4} = 663.3. \end{aligned}$$

2. Una vez que se ha realizado el suavizamiento por promedios móviles simples de toda la serie se emplea la ecuación 12.2 para calcular el segundo promedio móvil de cuatro trimestres:

$$\begin{aligned} M'_{16} &= \frac{M_{16} + M_{15} + M_{14} + M_{13}}{4} \\ &= \frac{663.3 + 655.6 + 648.8 + 627.3}{4} = 648.8. \end{aligned}$$

3. Se usa la igualdad 12.3 para calcular la diferencia de los dos promedios móviles:

$$\begin{aligned} a_{16} &= 2M_{16} - M'_{16} \\ &= 2 \times 663.3 - 648.8 = 677.8. \end{aligned}$$

   La ecuación 12.4 ajusta la pendiente:

$$\begin{aligned} b_{16} &= \frac{2}{n-1}\left(M_{16} - M'_{16}\right) \\ &= \frac{2}{4-1}(663.3 - 648.8) = 9.7. \end{aligned}$$

4. Realizamos el pronóstico de un periodo a futuro (ecuación 12.5 con $p = 1$):

$$\begin{aligned} \widehat{Y}_{16+1} &= a_{16} + b_{16}p \\ &= 677.8 + 9.7 \times 1 = 687.5. \end{aligned}$$

El pronóstico para el primer trimestre del año 2007 es de $\widehat{Y}_{17} = 687.5$, que puede parecer más razonable que los dos resultados presentados anteriormente.

12.3.3. Suavizamiento exponencial

El suavizamiento o atenuación exponencial es un método que se utiliza para revisar constantemente la estimación de los coeficientes de un modelo de pronósticos con base en cada observación sucesiva real.

El método se basa en promediar los valores anteriores de una serie, haciendo esto de forma decreciente, asignando mayor peso a las más recientes. Las ponderaciones empleadas se designan como α $(0 < \alpha < 1)$ para la observación más reciente, $\alpha(1-\alpha)$ para la siguiente más reciente, $\alpha(1-\alpha)^2$ para la siguiente, y así sucesivamente.

El método de cálculo de los pronósticos es el siguiente:

$$\text{nuevo pronóstico} = \alpha \times (\text{nueva observación}) + (1-\alpha) \times (\text{pronóstico anterior}),$$

de manera que la ecuación de la atenuación exponencial es

$$\widehat{Y}_{t+1} = \alpha Y_t + (1-\alpha)\widehat{Y}_t,$$

donde

$\widehat{Y}_{t+1}$ = nuevo valor atenuado o valor del pronóstico para el siguiente periodo;

α = constante de atenuación $(0 < \alpha < 1)$;

Y_t = nueva observación o valor real en el periodo t;

$\widehat{Y}_t$ = valor atenuado anterior o experiencia promedio de la serie atenuada al periodo $t-1$.

Una mejor interpretación de la ecuación que define el suavizamiento exponencial se puede ver en la siguiente descomposición:

$$\begin{aligned} \widehat{Y}_{t+1} &= \alpha Y_t + (1-\alpha)\widehat{Y}_t = \alpha Y_t + \widehat{Y}_t - \alpha\widehat{Y}_t \\ &= \widehat{Y}_t + \alpha\left(Y_t - \widehat{Y}_t\right). \end{aligned}$$

La atenuación exponencial es simplemente el pronóstico anterior $\left(\widehat{Y}_t\right)$ más α veces el error $\left(Y_t - \widehat{Y}_t\right)$, en el pronóstico anterior.

La constante de suavizamiento α sirve como el factor para ponderar. El valor de α determina el grado hasta el cual la observación más reciente influye en el valor del pronóstico. Cuando α es cercano a 1, en el nuevo pronóstico la observación más influyente será la más reciente. Inversamente, cuando α es cercano a 0, en el pronóstico influirán todas las observaciones de manera similar.

Un método para estimar α consiste en encontrar aquel valor que minimiza el error cuadrático medio (**ECM**):

$$\mathbf{ECM} = \frac{\sum_{t=1}^{n}\left(Y_t - \widehat{Y}_t\right)^2}{n},$$

para diferentes valores de α. Para generar pronósticos futuros se elige el valor de α que produce el error más pequeño.

En la siguiente tabla se presentan los cálculos de los valores de las ponderaciones para dos valores de α. En ella se puede observar cómo influyen los valores pasados en los pronósticos, en dependencia de α.

	$\alpha = 0.1$		$\alpha = 0.6$	
Periodo	**Cálculos**	**Ponderación**	**Cálculos**	**Ponderación**
t		0.100		0.600
$t-1$	0.9×0.1	0.090	0.4×0.6	0.240
$t-2$	$0.9 \times 0.9 \times 0.1$	0.081	$0.4 \times 0.4 \times 0.6$	0.096
$t-3$	$0.9 \times 0.9 \times 0.9 \times 0.1$	0.073	$0.4 \times 0.4 \times 0.4 \times 0.6$	0.038
Los restantes		0.656		0.026
	Total	1.000		1.000

Ejemplo. (Continuación) Realizar el pronóstico, mediante suavizamiento exponencial, de los datos del Cuadro 12.1.

Solución: Los resultados se presentan en el Cuadro 12.4 con valores de la constante de atenuación de 0.1 y 0.6.

t	Ventas Y_t	Valor suavizado $\widehat{Y}_t$ ($\alpha = 0.1$)	Error de pronostico e_i	Valor suavizado $\widehat{Y}_t$ ($\alpha = 0.6$)	Error de pronostico e_i
1	598	598.0		598.0	
2	390	598.0	-208.0	598.0	-208.0
3	267	577.2	-310.2	473.2	-206.2
4	573	546.2	26.8	349.5	223.5
5	588	548.9	39.1	483.6	104.4
6	425	552.8	-127.8	546.2	-121.2
7	371	540.0	-169.0	473.5	-102.5
8	609	523.1	85.9	412.0	197.0
9	777	531.7	245.3	530.2	246.8
10	532	556.2	-24.2	678.3	-146.3
11	433	553.8	-120.8	590.5	-157.5
12	689	541.7	147.3	496.0	193.0
13	855	556.4	298.6	611.8	243.2
14	618	586.3	31.7	757.7	-139.7
15	460	589.5	-129.5	673.9	-213.9
16	720	576.5	143.5	545.6	174.4
17		590.9		650.2	
		EMC=	27 902.7	**EMC=**	33 991.3

Cuadro 12.4: Valores atenuados exponencialmente.

La serie atenuada se calcula asignando inicialmente $\widehat{Y}_1 = 598$, que es el primer valor observado, a continuación los cálculos se realizan de la siguiente manera:

$$\begin{aligned} \widehat{Y}_{1+1} &= \alpha Y_1 + (1-\alpha)\widehat{Y}_1 \\ \widehat{Y}_2 &= (0.1)598 + (0.9)598 = 598. \end{aligned}$$

El error de pronóstico es

$$e_2 = Y_2 - \widehat{Y}_2 = 390 - 598 = -208.$$

Para el periodo 17, resulta

$$\begin{aligned}\widehat{Y}_{17} &= \alpha Y_{16} + (1-\alpha)\widehat{Y}_{16}\\ &= (0.1)720 + (0.9)576.5 = 590.9.\end{aligned}$$

Con similares cálculos, para una constante $\alpha = 0.6$ el pronóstico del periodo 17 es $\widehat{Y}_{17} = 650.2$.

Para definir cuál de los dos pronósticos es el más aceptable, se pueden comparar los errores cuadráticos medios de las series.

12.3.4. Suavizamiento exponencial doble

La técnica de suavizamiento exponencial doble, también conocida como *método de Brown*, se emplea para pronosticar series que tienen una tendencia lineal y es semejante al atenuamiento por la media móvil doble.

Primero, conviene que tengamos en cuenta que debido a que los valores de la serie no son pronósticos en sí mismos, las ecuaciones de actualización son más comprensibles si se adopta la siguiente notación.

$A_t =$ valor (simple) suavizado exponencialmente de Y_t en el periodo t.

$A'_t =$ valor doblemente suavizado exponencialmente de Y_t en el periodo t.

La técnica de suavizamiento exponencial doble se realiza de la forma siguiente:

1. Se calcula el valor simple suavizado exponencialmente, con el método antes expuesto:

$$A_{t+1} = \alpha Y_{t+1} + (1-\alpha)A_t. \tag{12.6}$$

2. Se calcula el valor doblemente suavizado exponencialmente, realizando otro suavizamiento exponencial a los datos recién obtenidos:

$$A'_{t+1} = \alpha A_{t+1} + (1-\alpha)A'_t. \tag{12.7}$$

3. Se calcula la diferencia de los valores atenuados exponencialmente:

$$a_t = 2A_t - A'_t. \tag{12.8}$$

y un factor de ajuste adicional, similar a la pendiente:

$$b_t = \frac{\alpha}{1-\alpha}\left(A_t - A'_t\right). \tag{12.9}$$

4. Se forma la ecuación para realizar el pronóstico de p periodos en el futuro:

$$\widehat{Y}_{t+p} = a_t + b_t p. \tag{12.10}$$

Ejemplo. (Continuación) Realizar el pronóstico, mediante suavizamiento exponencial doble, de los datos del Cuadro 12.1.

Solución: Los resultados de los cálculos se encuentran en el Cuadro 12.5.

A continuación se presentan los cálculos correspondientes para el pronóstico del periodo 17.

t	Ventas Y_t	Valor suavizado A_t	Valor suavizado A'_t	Valor de a	Valor de b	Pronostico $a+bp$ $(p=1)$
1	598	598.0	598.0	598.0	0.0	
2	390	577.2	595.9	558.5	-2.1	598.0
3	267	546.2	590.9	501.4	-5.0	556.4
4	573	548.9	586.7	511.0	-4.2	496.4
5	588	552.8	583.3	522.2	-3.4	506.8
6	425	540.0	579.0	501.0	-4.3	518.8
7	371	523.1	573.4	472.8	-5.6	496.7
8	609	531.7	569.2	494.1	-4.2	467.2
9	777	556.2	567.9	544.5	-1.3	490.0
10	532	553.8	566.5	541.1	-1.4	543.2
11	433	541.7	564.0	519.4	-2.5	539.7
12	689	556.4	563.3	549.6	-0.8	516.9
13	855	586.3	565.6	607.0	2.3	548.8
14	618	589.5	568.0	611.0	2.4	609.3
15	460	576.5	568.8	584.2	0.9	613.4
16	720	590.9	571.0	610.7	2.2	585.1
17						612.9

Cuadro 12.5: Pronóstico por suavizamiento exponencial doble con $\alpha = 0.1$.

1. Mediante la ecuación 12.6 se calcula el valor de la atenuación exponencial simple de la serie inicial (columna 3):

$$\begin{aligned} A_{16} &= \alpha Y_{16} + (1-\alpha)A_{15} \\ &= (0.1)720 + (0.9)576.5 = 590.9. \end{aligned}$$

2. La ecuación 12.7 se usa para calcular el valor doblemente suavizado exponencialmente (columna 4).

$$\begin{aligned} A'_{16} &= \alpha A_{16} + (1-\alpha)A'_{15} \\ &= (0.1)590.9 + (0.9)568.8 = 571.0. \end{aligned}$$

3. En la columna 5 se calculan las diferencias entre los valores atenuados exponencialmente, mediante la igualdad 12.8:

$$\begin{aligned} a_{16} &= 2A_{16} - A'_{16} \\ &= 2(590.9) - 571.0 = 610.7. \end{aligned}$$

4. Con la ecuación 12.9 se calcula el valor adicional de ajuste (columna 6).

$$\begin{aligned} b_{16} &= \frac{\alpha}{1-\alpha}\left(A_{16} - A'_{16}\right) \\ &= \frac{0.1}{0.9}(590.9 - 571.0) = 2.2. \end{aligned}$$

5. Finalmente, se emplea la igualdad 12.10 para hacer el pronóstico en el futuro:

$$\begin{aligned} \widehat{Y}_{16+1} &= a_{16} + b_{16}p \\ &= 610.7 + 2.2(1) = 612.9. \end{aligned}$$

12.3.5. Método de Holt

El método de Holt es una técnica que se emplea para manejar series de datos que tienen una tendencia lineal pero que no presentan estacionalidad. La técnica atenúa en forma directa la tendencia y la pendiente utilizando diferentes constantes de suavizamiento para cada una de ellas. Por esta razón, el método presenta mayor flexibilidad al seleccionar los modelos.

La técnica consiste de los siguientes pasos:

1. Se suaviza la serie mediante el método exponencial:

$$A_t = \alpha Y_t + (1-\alpha)\left(A_{t-1} + T_{t-1}\right). \tag{12.11}$$

2. Se estima la tendencia:

$$T_t = \beta\left(A_t - A_{t-1}\right) + (1-\beta)T_{t-1}. \tag{12.12}$$

3. Se realiza el pronóstico de p periodos en el futuro:

$$\widehat{Y}_{t+p} = A_t + pT_t, \tag{12.13}$$

donde

$A_t =$ nuevo valor atenuado;

$\alpha =$ constante de atenuación exponencial de los datos $(0 < \alpha < 1)$;

$Y_t =$ nueva observación o valor real en el periodo t;

$\beta =$ constante de atenuación para la estimación de la tendencia $(0 < \beta < 1)$;

$T_t =$ estimación de la tendencia;

$\widehat{Y}_{t+p} =$ pronóstico de p periodos en el futuro.

La primera ecuación es similar a la igualdad original de atenuación exponencial simple, con excepción de que se incorpora el término (T_t) para la tendencia.

En la segunda ecuación se toman dos valores sucesivos de atenuación exponencial, ya que ellos se atenuaron con fines de eliminar la aleatoriedad, su diferencia constituye una estimación de la tendencia local de los datos. Se atenúa esta tendencia con la constante β y se le suma la tendencia anteriormente calculada multiplicada por $1-\beta$. El valor obtenido es una tendencia atenuada que excluye cualquier aleatoriedad.

La ecuación del pronóstico suma el nivel actual de los datos A_t y el producto de la tendencia T_t por los p periodos.

Notemos que para iniciar la iteración, se toma $A_1 = Y_1$ y $T_1 = 0$.

Ejemplo. (Continuación) Realizar el pronóstico, con el método de Holt, de los datos del Cuadro 12.1.

Solución: Los resultados de los cálculos se encuentran en el Cuadro 12.6.

Examinemos cómo se realiza el pronóstico en los periodos 2 y 17 con $\alpha = 0.3$ y $\beta = 0.2$.

t	Ventas Y_t	Valor de la actualización A_t	Valor de la tendencia T_t	Pronóstico $a+bp$ $\widehat{Y}_t$
1	598	598.0	0.0	
2	390	535.6	-12.5	
3	267	446.3	-27.8	523.1
4	573	464.8	-18.6	418.4
5	588	488.8	-10.1	446.2
6	425	462.6	-13.3	478.7
7	371	425.8	-18.0	449.3
8	609	468.2	-5.9	407.8
9	777	556.7	13.0	462.3
10	532	558.4	10.7	569.6
11	433	528.2	2.5	569.1
12	689	578.3	12.0	530.8
13	855	669.7	27.9	590.3
14	618	673.7	23.1	697.6
15	460	625.8	8.9	696.9
16	720	660.3	14.0	634.7
17				674.3

Cuadro 12.6: Pronóstico mediante el método de Holt.

1. Cálculo del valor de la actualización de la serie. Para iniciar el proceso tomamos $A_1 = Y_1$ y $T_1 = 0$:

$$\begin{aligned} A_2 &= \alpha Y_2 + (1-\alpha)\left(A_{2-1} + T_{2-1}\right) \\ &= 0.3(390) + (1-0.3)(598+0) = 535.6. \end{aligned}$$

2. La estimación de la tendencia:

$$\begin{aligned} T_2 &= \beta\left(A_2 - A_{2-1}\right) + (1-\beta)T_{2-1} \\ &= 0.2(535.5-598) + (1-0.2)0 = -12.5. \end{aligned}$$

3. El pronóstico de un periodo en el futuro:

$$\begin{aligned} \widehat{Y}_{2+1} &= A_2 + pT_2 \\ &= 535.6 + (1)(-12.5) = 523.1. \end{aligned}$$

Con el mismo proceso para el periodo 17 tenemos:

1. Actualización de la serie:

$$\begin{aligned} A_{16} &= \alpha Y_{16} + (1-\alpha)\left(A_{16-1} + T_{16-1}\right) \\ &= 0.3(720) + (1-0.3)(625.8+8.9) = 660.3. \end{aligned}$$

2. Estimación de la tendencia:

$$\begin{aligned} T_{16} &= \beta\left(A_{16} - A_{16-1}\right) + (1-\beta)T_{16-1} \\ &= 0.2(660.3-625.8) + (1-0.2)8.9 = 14.0. \end{aligned}$$

3. Pronóstico de un periodo en el futuro:

$$\begin{aligned} \widehat{Y}_{16+1} &= A_{16} + T_{16}\,p \\ &= 660.3 + (14.0)(1) = 674.3. \end{aligned}$$

12.3.6. Método de Winters

La atenuación por el método de Winters es apropiada cuando la serie presenta una tendencia lineal y una variación estacional. Es una extensión del modelo de Holt, en la que se utiliza una ecuación adicional para estimar la estacionalidad mediante un índice estacional.

Los pasos que se siguen en este método son:

1. Se suaviza la serie exponencialmente:

$$A_t = \alpha \frac{Y_t}{S_{t-L}} + (1-\alpha)(A_{t-1} + T_{t-1}). \tag{12.14}$$

2. Se estima la tendencia:

$$T_t = \beta (A_t - A_{t-1}) + (1-\beta) T_{t-1}. \tag{12.15}$$

3. Se estima la estacionalidad:

$$S_t = \gamma \frac{Y_t}{A_t} + (1-\gamma) S_{t-L}. \tag{12.16}$$

4. Se realiza el pronóstico de p periodos en el futuro:

$$\widehat{Y}_{t+p} = (A_t + T_t\, p)\, S_{t-L+p}, \tag{12.17}$$

donde

$A_t =$ nuevo valor atenuado;

$\alpha =$ constante de atenuación $(0 < \alpha < 1)$;

$Y_t =$ nueva observación o valor real en el periodo t;

$\beta =$ constante de atenuación de la estimación de la tendencia $(0 < \beta < 1)$;

$T_t =$ estimación de la tendencia;

$\gamma =$ constante de atenuación de la estimación de la estacionalidad $(0 < \gamma < 1)$;

$S_t =$ estimación de la estacionalidad;

$L =$ longitud de la estacionalidad;

$\widehat{Y}_{t+p} =$ pronóstico de p periodos en el futuro.

Obsérvese que el índice estacional S_t se calcula tomando en cuenta el índice estacional S_{t-L}, correspondiente a L periodos anteriores, y que para iniciar la iteración se toma $A_1 = Y_1$, $T_1 = 0$ y $S_k = 1$, cuando el índice k no es positivo.

Ejemplo. (Continuación) Realizar el pronóstico, mediante el método de Winters, de los datos del Cuadro 12.1.

Solución: Los resultados de la aplicación de la técnica se encuentran en el Cuadro 12.7.

Para comenzar los cálculos, vemos que la longitud de la estacionalidad es $L = 4$ y que se requieren las estimaciones iniciales del valor atenuado, de la tendencia y cuatro estimaciones de la estacionalidad. Para principiar el proceso de cálculo, usaremos 598 como valor inicial de la actualización, 0 como el valor de la estimación inicial de la tendencia y 1 como la estimación de la estacionalidad.

Analicemos los cálculos de los periodos 2 y 17, utilizando los valores de $\alpha = 0.3$, $\beta = 0.2$ y $\gamma = 0.1$.

t	Ventas Y_t	Valor de la actualizacion A_t	Valor de la tendencia T_t	Valor de la estacionalidad S_t	Pronostico $a+bp$ Y_t
1	598	598.0	0.0	1.00	598.0
2	390	535.6	-12.5	0.97	598.0
3	267	344.4	-48.2	0.98	598.0
4	573	325.0	-42.5	1.08	598.0
5	588	547.8	10.6	1.01	282.6
6	425	550.1	8.9	0.95	543.2
7	371	417.6	-19.3	0.97	546.4
8	609	415.9	-15.8	1.12	428.6
9	777	646.6	33.5	1.03	403.0
10	532	734.8	44.4	0.93	648.0
11	433	537.6	-3.9	0.95	754.8
12	689	485.7	-13.5	1.15	595.1
13	855	722.7	36.6	1.04	484.9
14	618	823.5	49.4	0.91	706.0
15	460	612.1	-2.7	0.93	831.3
16	720	508.7	-22.9	1.17	698.0
17					637.2
18					556.0
19					566.7
20					698.8

Cuadro 12.7: Pronóstico mediante el método de Winters.

1. Cálculo del valor de la actualización de la serie. Para iniciar el proceso tomamos $A_1 = Y_1$, $T_1 = 0$ y $S_{2-4} = 1$:

$$\begin{aligned} A_2 &= \alpha \frac{Y_2}{S_{2-4}} + (1-\alpha)(A_{2-1} + T_{2-1}) \\ &= 0.3\frac{390}{1.0} + (1-0.3)(598+0) = 535.6. \end{aligned}$$

2. La estimación de la tendencia:

$$\begin{aligned} T_2 &= \beta(A_2 - A_{2-1}) + (1-\beta)T_{2-1} \\ &= 0.2(535.5 - 598) + (1-0.2)0 = -12.5. \end{aligned}$$

3. La estimación de la estacionalidad:

$$\begin{aligned} S_2 &= \gamma\frac{Y_2}{A_2} + (1-\gamma)S_{2-4} \\ &= 0.1\frac{390}{535.5} + (1-0.1)1 = 0.97. \end{aligned}$$

De manera semejante, tenemos para el periodo 16 y los sucesivos pronósticos:

1. Actualización de la serie:

$$\begin{aligned} A_{16} &= \alpha \frac{Y_{16}}{S_{16-4}} + (1-\alpha)(A_{16-1} + T_{16-1}) \\ &= 0.3\frac{720}{1.15} + (1-0.3)(612.1 + (-2.7)) = 614.41. \end{aligned}$$

2. Estimación de la tendencia:

$$\begin{aligned} T_{16} &= \beta(A_{16} - A_{16-1}) + (1-\beta)T_{16-1} \\ &= 0.2(614.41 - 612.1) + (1-0.2)(-2.7) = -1.698. \end{aligned}$$

3. La estimación de la estacionalidad:

$$\begin{aligned} S_{16} &= \gamma \frac{Y_{16}}{A_{16}} + (1-\gamma) S_{16-4} \\ &= 0.1 \frac{720}{614.41} + (1-0.1)(1.15) = 1.15. \end{aligned}$$

4. Pronósticos:

a) De un periodo en el futuro, $p = 1$:

$$\begin{aligned} \widehat{Y}_{16+1} &= (A_{16} + T_{16}\,p)\, S_{16-4+1} \\ &= [614.41 + (-1.698)(1)]\,(1.04) = 637.2. \end{aligned}$$

b) De dos periodos en el futuro, $p = 2$:

$$\begin{aligned} \widehat{Y}_{16+2} &= (A_{16} + T_{16}\,p)\, S_{16-4+2} \\ &= [614.41 + (-1.698)(2)]\,(0.91) = 556.0. \end{aligned}$$

c) De tres periodos en el futuro, $p = 3$:

$$\begin{aligned} \widehat{Y}_{16+3} &= (A_{16} + T_{16}\,p)\, S_{16-4+3} \\ &= [614.41 + (-1.698)(3)]\,(0.93) = 566.7. \end{aligned}$$

d) De dos periodos en el futuro, $p = 4$:

$$\begin{aligned} \widehat{Y}_{16+4} &= (A_{16} + T_{16}\,p)\, S_{16-4+4} \\ &= [614.41 + (-1.698)(4)]\,(1.15) = 698.8. \end{aligned}$$ ◀

El método de Winters presenta la ventaja sobre los otros métodos en que podemos realizar pronósticos por más periodos (por lo menos por un ciclo completo de estacionalidad).

Hemos expuesto varios métodos sencillos de pronosticación de series de tiempo. En este punto surge la siguiente inquietud: ¿cuál método emplear?

12.4. Comparación de los métodos

Primero tengamos en cuenta que no existe un método que sea el mejor en toda ocasión. La elección del método dependerá del conjunto de datos que se disponga y de la finalidad de los mismos.

La comparación de las técnicas de atenuación se realiza mediante el error cuadrático medio que, como ya se indicó, se calcula por

$$\mathbf{ECM} = \frac{\sum_{t=1}^{n} \left(Y_t - \widehat{Y}_t\right)^2}{n} = \frac{\sum_{t=1}^{n} e_t^2}{n},$$

donde e_t es el error de pronóstico para cada una de las observaciones de la serie. Entonces, podemos enunciar la siguiente regla:

Aquel método que proporcione el menor **ECM** será el más apropiado para la serie de tiempo que se analiza.

Ejemplo. (Continuación) Para los datos del Cuadro 12.1 veamos cuál es el método más apropiado.

Solución: En el siguiente cuadro se encuentra el resumen de los métodos analizados con su **ECM**.

Método	Parámetros	ECM
Promedio móvil simple		20 826.2
Promedio móvil doble		24 834.8
Suavizamiento exponencial	$\alpha = 0.1$	27 902.7
Suavizamiento exponencial	$\alpha = 0.6$	33 991.3
Suavizamiento exponencial doble	$\alpha = 0.1$	29 547.4
Método de Holt	$\alpha = 0.1,\ \beta = 0.2$	31 634.7
Método de Winters	$\alpha = 0.1,\ \beta = 0.2,\ \gamma = 0.1$	54 548.9

De los resultados, aparentemente, el mejor método es el promedio móvil simple y el pronóstico basado en él sería el apropiado.[1] ◀

Dependiendo de las observaciones se sugiere que se utilicen los modelos:

Exponencial. Si la serie no tiene una tendencia marcada y no muestra variación estacional.

Holt. Si la serie tiene tendencia lineal pero no muestra estacionalidad.

Winters. Si la serie tiene tendencia lineal y muestra variación estacional.

Respecto a los parámetros de suavizamiento debemos indicar que ellos siempre varían entre 0 y 1. Para cada constante de atenuación, cuanto mayor sea su valor, tanto mayor importancia se le dará a la observación más reciente. Téngase presente que la mayoría de los programas estadísticos que disponen de estos métodos tienen implementados algoritmos que escogen automáticamente la combinación de parámetros que dan el **ECM** mínimo, lo que permite tener una estimación apropiada a las observaciones disponibles.

Se aconseja al lector que implemente las fórmulas de los diversos métodos en una hoja electrónica, lo que le facilitará la realización de los cálculos.

12.5. Ejercicios

1. En una academia de idiomas se sigue un sistema trimestral. El número de alumnos que estudian el idioma esperanto en cada trimestre, durante 4 años, se resume en la siguiente tabla:

Año	Trimestre	No. alumnos	Año	Trimestre	No. alumnos
1	1	10	3	1	13
	2	31		2	34
	3	43		3	48
	4	16		4	19
2	1	11	4	1	15
	2	33		2	37
	3	45		3	51
	4	17		4	21

a) Grafique los datos;

Mediante los métodos que se indican, realice el suavizamiento de la serie y el pronóstico al periodo señalado. También, calcule el **ECM** respectivo.

[1] Como el objetivo del ejemplo es ilustrar el funcionamiento de las técnicas, los valores de los parámetros de suavizamiento se tomaron de manera arbitraria; ellos no son los óptimos para ninguno de los métodos.

b) promedios móviles de orden 4 y $p = 1$;

c) media móvil doble de orden 4 y $p = 1$;

d) suavizamiento exponencial con $\alpha = 0.2$ y $p = 1$;

e) suavizamiento exponencial doble con $\alpha = 0.2$ y $p = 1$;

f) método de Holt con $\alpha = 0.2$, $\beta = 0.3$;

g) método de Winters para cuatro trimestres, con $\alpha = 0.1$, $\beta = 0.3$ y $\gamma = 0.4$.

2. Las ventas trimestrales de casas que ha realizado una compañía inmobiliaria en los últimos años se presenta a continuación:

Año	Trimestre	Ventas	Año	Trimestre	Ventas
1	1	50	4	1	55
	2	35		2	35
	3	25		3	25
	4	40		4	55
2	1	45	5	1	55
	2	35		2	40
	3	20		3	35
	4	30		4	60
3	1	35	6	1	75
	2	20		2	50
	3	15		3	40
	4	40		4	65

a) Grafique los datos;

Mediante los métodos indicados, realice el suavizamiento de la serie y el pronóstico al periodo señalado.

b) promedios móviles de orden 4 y $p = 1$;

c) media móvil doble de orden 4 y $p = 1$;

d) suavizamiento exponencial con $\alpha = 0.1$ y $p = 1$;

e) suavizamiento exponencial doble con $\alpha = 0.3$;

f) método de Holt con $\alpha = 0.3$, $\beta = 0.25$;

g) método de Winters para cuatro trimestres, con $\alpha = 0.4$, $\beta = 0.1$ y $\gamma = 0.3$;

h) Encuentre el mejor método de ajuste mediante el criterio **ECM**.

3. En una universidad se implantó el sistema de estudios cuatrimestral. El número de alumnos matriculados en la materia de geometría se presenta a continuación:

Periodo	Alumnos	Periodo	Alumnos
1	54	9	101
2	58	10	103
3	65	11	102
4	72	12	110
5	73	13	112
6	71	14	111
7	93	15	128
8	94		

a) Grafique los datos;

Mediante los métodos indicados, realice el suavizamiento de la serie y el pronóstico al periodo señalado.

b) promedios móviles de orden 3 y $p = 1$;

c) media móvil doble de orden 3 y $p = 1$;

d) suavizamiento exponencial con $\alpha = 0.4$ y $p = 1$;

e) suavizamiento exponencial doble con $\alpha = 0.4$ y $p = 1$;

f) método de Holt con $\alpha = 0.35$, $\beta = 0.15$;

g) método de Winters para tres trimestres, con $\alpha = 0.2$, $\beta = 0.15$ y $\gamma = 0.1$;

h) Encuentre el mejor método de ajuste mediante el criterio **ECM**.

Capítulo 13

Elementos de Muestreo

Por un pequeño pedazo de muestra podemos juzgar a la pieza completa
Miguel de Cervantes Saavedra

Como se indicó en el Capítulo 1, en muchos estudios estadísticos la recolección de la información se realiza mediante la investigación por muestreo y se sacan conclusiones con base en la «inferencia estadística».

Este capítulo está dedicado a exponer los métodos básicos empleados en las investigaciones por muestreo, pero no solo en lo que tiene que ver con la parte estadística, también se dan algunas indicaciones para la realización práctica de los sondeos.

13.1. Conceptos básicos

En las investigaciones cuyo objetivo es la recopilación de información estadística sobre un grupo de personas o cosas, se distinguen dos tipos de estudios:

1. La encuesta total o censo, cuyo objeto es examinar a todos los elementos de la población.
2. La encuesta parcial o por muestreo, que tiene por objeto examinar una parte pequeña de la población, e inferir resultados para aplicarlos a la población completa.

Aquí aparecen dos conceptos que frecuentemente utilizaremos: la población (o universo muestral) y la muestra.

Definición (de población) Una población es una colección completa de personas, animales, plantas o cosas de las cuales se desea recolectar datos.

Definición (de muestra) Es un grupo de unidades seleccionadas de un grupo mayor (la población).

Por el estudio de la muestra se espera obtener conclusiones sobre la población.

En muchos casos la elección de una muestra es fácil; por ejemplo, para conocer la proporción de desperdicio en un saco de maíz o para conocer el número de piezas defectuosas en un lote de productos,

etc. Bastará mezclar bien los productos y tomar una pequeña parte de ellos. Lo que se deduzca de esta pequeña porción (o muestra) servirá para juzgar a la totalidad del producto.

Para conocer la preferencia de una marca de gaseosa por parte de los consumidores, o el tiempo que la población dedica a mirar la televisión o el favoritismo por cierto partido político entre los votantes, es más complicado. Los elementos son más heterogéneos y sería imposible aplicar el método descrito para tomar una muestra, como se hace con los productos.

Aunque la muestra podría tomarse como un subconjunto cualquiera de la población, en una investigación es necesario que cumpla con ciertos requisitos, para que nos provea de información confiable sobre la población. La muestra ha de ser una reproducción en pequeño de la población. Así, ha de ser un Ecuador, un Guayas o un Quito en miniatura, si se trata de un sondeo relativo al Ecuador, o la provincia del Guayas o a la ciudad de Quito.

Pero, ¿cómo hacer para que la muestra sea una copia, en pequeño, de la población? La respuesta es que la muestra debe estar constituida por un número suficiente de elementos, tomados al azar, de la población.

Para la correcta elección de la muestra, en primer lugar es necesario hacer una lista de los objetos de la cual se la seleccionará, luego se procederá a sortearlos para incluirlos en la muestra.

Entonces, tenemos las siguientes definiciones:

Definición (de unidad muestral) Los objetos que se seleccionan de una población se llaman unidades muestrales.

Definición (de marco muestral) Un marco muestral es una lista completa de todas las unidades muestrales de la población.

Por ejemplo, se desea conocer las preferencias electorales de todos los miembros adultos de la ciudad de Ambato. La población está constituida por todas las personas en capacidad de votar que viven en Ambato. El marco muestral es una lista completa con los nombres de cada miembro de la población (el padrón electoral). Una unidad muestral es un residente en Ambato y que esté en capacidad de votar.

Observemos que no siempre es posible tener un marco muestral perfectamente definido, ya sea porque éste es muy grande, o no existe, o no se lo puede confeccionar. Más aún, solo cuando la población es pequeña o controlable por el investigador, es posible contar con un marco muestral ideal. Preguntémonos: ¿quién podría elaborar una lista de todos los individuos, o de todas las familias, que viven en Guayaquil?

Antes de terminar, tenemos que referirnos a las conveniencias y a las limitaciones de las muestras, con respecto a la realización de un censo.

Ventajas de tomar muestras:

1. Son más económicas.
2. Se emplea menor tiempo en la recolección de los datos.
3. Se obtiene una mejor calidad de información.
4. Se emplea cuando la población es grande o puede considerarse infinita.
5. Son apropiadas cuando el proceso de medida de cada elemento es destructivo o conlleva riesgos a la salud.

Limitaciones de tomar muestras:

1. Si se necesita información de todos los elementos que conforman el universo estadístico.
2. Si se requiere información muy desagregada, para áreas muy pequeñas.
3. Cuando no existen los elementos técnicos y humanos que garanticen un buen diseño muestral y una buena ejecución del sondeo.

A continuación expondremos los principales tipos de investigaciones por muestreo, que son el aleatorio simple, el estratificado y el de conglomerados.

13.2. Muestreo aleatorio simple

Como se dijo, un buen diseño muestral requiere que los elementos escogidos sean tomados al azar. Con ésto garantizamos que la muestra represente a la población y que las inferencias a realizar sean válidas. Al examen de este tipo de investigación dedicaremos esta sección.

La mayoría de sondeos tiene uno de los tres objetivos siguientes: estimar el total poblacional τ, o estimar la media de una población μ, o estimar la proporción poblacional p.

13.2.1. Estimación del total poblacional

Si una población está constituida por N unidades, de las cuales interesa medir el caracter x de cada uno: x_1, x_2, ..., x_N. El total poblacional se define por

$$\tau = \sum_{i=1}^{N} x_i = N\mu,$$

donde μ es la media poblacional.

Por ejemplo, en una encuesta realizada para estimar los gastos en salud de los habitantes de una ciudad, se investigaría el gasto medio por persona, μ, que realizan en un año; o también, puede ser de interés el gasto total, τ, que se realiza en dicho poblado por concepto de salud.

Intervalo de confianza

El intervalo para el total poblacional τ, a un nivel de confianza del $(1-\alpha)\times 100\,\%$ es

$$\left(\widehat{\tau} - z_{\alpha/2}\frac{Ns}{\sqrt{n}}\sqrt{\frac{N-n}{N}};\ \widehat{\tau} + z_{\alpha/2}\frac{Ns}{\sqrt{n}}\sqrt{\frac{N-n}{N}}\right).$$

Donde:

N es el número de elementos en la población.

n es el número de elementos en la muestra.

$\widehat{\tau}$ es la estimación del total poblacional: $\widehat{\tau} = N\overline{x}$.

$\overline{x}$ es el promedio de la muestra.

s es la desviación estándar de la muestra.

$z_{\alpha/2}$ el coeficiente de la ley normal estándar[1] para el cual el área en el extremo superior es igual a $\dfrac{\alpha}{2}$.

Tamaño de la muestra

La cantidad que hay que sumarle o restarle a un estimador, en la confección del intervalo de confianza, se denomina error. En nuestro caso

$$E_\tau = z_{\alpha/2}\frac{Ns}{\sqrt{n}}\sqrt{\frac{N-n}{N}}.$$

De aquí, si se desea tener una estimación al nivel $(1-\alpha)\times 100\,\%$ de confianza, con un error E_τ dado, a partir de una muestra obtenida de una población de tamaño N, el número de unidades a incluir en el sondeo es

$$n = \frac{\left(z_{\alpha/2}Ns\right)^2}{E_\tau^2 + \left(z_{\alpha/2}\right)^2 Ns^2}.$$



Ejemplo. Una empresa de telefonía celular desea estimar el tiempo total que se emplean sus líneas en un fin de semana. Se seleccionó al azar una muestra de 420 clientes, de los 62 000 que habían hecho uso de sus teléfonos y se registró el tiempo de uso. El tiempo promedio y la desviación estándar de la muestra fueron $\overline{x} = 3.61$ min y $s = 1.28$ min. A un nivel del 95.5 %: a) obtener un intervalo de confianza para el tiempo total de uso de los teléfonos ese fin de semana; b) Considerando una desviación estándar de 1.25, calcular el tamaño de la muestra para que el error sea menor o igual a 20 000 minutos.

Solución:

a) Para este ejemplo, $N = 62\,000$, $n = 420$, $\overline{x} = 3.61$, $s = 1.28$ y $z_{0.0225} = 2$. Entonces,

$$\widehat{\tau} = N\overline{x} = 62\,000 \times 3.61 = 223\,820.$$

El intervalo es

$$\left(\widehat{\tau} - z_{\alpha/2}\frac{Ns}{\sqrt{n}}\sqrt{\frac{N-n}{N}};\widehat{\tau} + z_{\alpha/2}\frac{Ns}{\sqrt{n}}\sqrt{\frac{N-n}{N}}\right)$$

$$\left(223\,820 - 2\frac{62\,000\times 1.28}{\sqrt{420}}\sqrt{\frac{62\,000-420}{62\,000}}; 223\,820 + 2\frac{62\,000\times 1.28}{\sqrt{420}}\sqrt{\frac{62\,000-420}{62\,000}}\right),$$

o sea (216 100; 231 540).

Por lo tanto, se estima que el tiempo total de uso de los teléfonos está entre los 216 mil y 231 mil minutos.

b) Con $s = 1.25$ y $E_\tau = 20\,000$, se tiene:

$$\begin{aligned} n &= \frac{4N^2s^2}{E_\tau^2 + 4Ns^2} \\ &= \frac{4(62\,000)^2(1.25)^2}{(20\,000)^2 + 4(62\,000)(1.25)^2} = 60. \end{aligned}$$

Se necesitará consultar al menos a 60 clientes.

[1] En la práctica se utilizan valores de $z_{\alpha/2}$ iguales a 2, cuando se trabaja con una confiabilidad del 95.5 %, o a 3, cuando la confiabilidad es del 99.7 %.

13.2.2. Estimación de la media poblacional

Como se dijo, puede ser necesario estimar la media poblacional, mediante la información obtenida en un sondeo; por ejemplo, cuando se afirma que el nivel medio de escolaridad en el Ecuador es de 7.5 años, o si se desea conocer el consumo medio anual de café de la población adulta en una localidad.

Intervalo de confianza

El intervalo para la media poblacional μ a un nivel de confianza del $(1-\alpha)\times 100\,\%$ es

$$\left(\overline{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N}}; \overline{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N}}\right).$$

Tamaño de la muestra

El tamaño de la muestra necesaria para tener un error prefijado E_μ, a un nivel de confianza de nivel $(1-\alpha)\times 100\,\%$, a partir de una población de tamaño N es:

$$n = \frac{(z_{\alpha/2})^2 N s^2}{N E_\mu^2 + (z_{\alpha/2})^2 s^2}.$$



Ejemplo. En un estudio médico sobre el consumo de tabaco, por la población adulta, en una ciudad de un millón de habitantes adultos, se consultó a 120 personas. Los resultados de la investigación mostraron un consumo promedio diario de 3.8 cigarrillos, por persona, con una desviación estándar de 1.1. a) Determinar el intervalo al 97 % para el número promedio de cigarrillos que se consumen; b) ¿A cuántos individuos ha de consultarse para que la estimación del número medio de cigarrillos quede a menos de 0.3 del valor verdadero?, si se considera un nivel de confianza del 95 %.

Solución:

a) Tenemos que $N = 1\,000\,000$, $n = 120$, $\overline{x} = 3.8$, $s = 1.1$ y $z_{0.015} = 2.17$.

El intervalo es

$$\left(\overline{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N}}; \overline{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N}}\right)$$
$$\left(3.8 - 2.17\frac{1.1}{\sqrt{120}}\sqrt{\frac{1\,000\,000-120}{1\,000\,000}}; 3.8 + 2.17\frac{1.1}{\sqrt{120}}\sqrt{\frac{1\,000\,000-120}{1\,000\,000}}\right)$$
$$(3.8 - 0.22; 3.8 + 0.22)$$

Entonces, el intervalo es (3.58; 4.02).

b) Con los datos previos y $E_\mu = 0.3$ y $z_{0.025} = 1.96$:

$$\begin{aligned} n &= \frac{\left(z_{\alpha/2}\right)^2 N s^2}{N E_\mu^2 + \left(z_{\alpha/2}\right)^2 s^2} \\ &= \frac{(1.96)^2(1\,000\,000)(1.1)^2}{(1\,000\,000)(0.3)^2 + (1.96)^2(1.1)^2} = 51.6. \end{aligned}$$

Por lo tanto, se ha de investigar al menos a 52 personas, para tener el error deseado.

13.2.3. Estimación de la proporción poblacional

Este caso es, probablemente, el más empleado en las investigaciones de mercado y en los sondeos políticos; así, frecuentemente se encuentra en periódicos y revistas datos como éstos: el 70 % de la población rechaza la decisión del gobierno de aumentar el precio de los pasajes, o un 45 % de los consumidores de gaseosas prefiere una marca determinada.

Intervalo de confianza

El intervalo para la proporción p a un nivel de confianza del $(1-\alpha)\times 100\,\%$ es

$$\left(\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}\widehat{q}}{n-1}}\sqrt{\frac{N-n}{N}}; \widehat{p} + z_{\alpha/2}\sqrt{\frac{\widehat{p}\widehat{q}}{n-1}}\sqrt{\frac{N-n}{N}}\right).$$

Donde:

N es el número de elementos en la población.

n es el número de elementos en la muestra.

$\widehat{p}$ es la estimación de la proporción, que se calcula por $\widehat{p} = \frac{y}{n}$, siendo y es el número de individuos que responde favorablemente en la encuesta y $\widehat{q} = 1 - \widehat{p}$.

Tamaño de la muestra

El tamaño de la muestra necesaria para tener un error prefijado E_p, a un nivel de confianza del $(1-\alpha)\times 100\,\%$, a partir de una población de tamaño N es:

$$n = \frac{(z_{\alpha/2})^2 N\widehat{p}\widehat{q}}{NE_p^2 + (z_{\alpha/2})^2\widehat{p}\widehat{q}}.$$

Como muchas veces se desconoce la estimación $\widehat{p}$, se toma el tamaño máximo de la muestra, que se obtiene haciendo $\widehat{p} = \widehat{q} = 0.5$; entonces,

$$n = \frac{(z_{\alpha/2})^2 N}{4NE_p^2 + (z_{\alpha/2})^2}.$$

Ejemplos

1. Una empresa productora de comida deshidratada desea introducir al mercado su producto Sopainstant. Se realiza un estudio para determinar la proporción de hogares que preparan al menos una vez al mes sopa deshidratada. La muestra fue de tamaño 240 en una ciudad de 20 mil hogares, resultando 40 respuestas afirmativas. Si se considera un nivel de confianza del 98 %: a) Encontrar el intervalo de confianza para la proporción; b) Utilizando la proporción previamente calculada, ¿a cuántos hogares se ha de encuestar, si se desea tener un error de 0.06 en la estimación?; c) Calcular el tamaño máximo de muestra.

   Solución:

a) Se tiene que $N = 20\,000$, $n = 240$, $y = 40$ y $z_{0.01} = 2.33$. Entonces,

$$\widehat{p} = \frac{y}{n} = \frac{40}{240} = 0.167.$$

El intervalo es

$$\left(\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}\widehat{q}}{n-1}}\sqrt{\frac{N-n}{N}}; \widehat{p} + z_{\alpha/2}\sqrt{\frac{\widehat{p}\widehat{q}}{n-1}}\sqrt{\frac{N-n}{N}}\right)$$

$$\left(0.167 - 2.33\sqrt{\frac{0.167 \times 0.833}{239}}\sqrt{\frac{20\,000-240}{20\,000}}; 0.167 + 2.33\sqrt{\frac{0.167 \times 0.833}{239}}\sqrt{\frac{20\,000-240}{20\,000}}\right)$$

$$(0.111; 0.223).$$

Es decir, con un 98 % de confianza, la proporción de potenciales compradores del producto está entre el 11.1 % y el 22.3 % de los hogares de la localidad.

b) Consideremos que $\widehat{p} = 0.167$ y $E_p = 0.06$.

$$\begin{aligned} n &= \frac{(z_{\alpha/2})^2 N\widehat{p}\widehat{q}}{NE_p^2 + (z_{\alpha/2})^2\widehat{p}\widehat{q}} \\ &= \frac{(2.33)^2(20\,000)(0.167)(0.833)}{(0.06)^2(20\,000) + (2.33)^2(0.167)(0.833)} = 207.6. \end{aligned}$$

Lo que quiere decir que se deberán encuestar al menos 208 hogares.

c) Si suponemos que $\widehat{p}$ es desconocido y tomamos $\widehat{p} = 0.5$, obtenemos:

$$\begin{aligned} n &= \frac{(z_{\alpha/2})^2 N}{4NE_p^2 + (z_{\alpha/2})^2} \\ &= \frac{(2.33)^2 20\,000}{4(20\,000)(0.06)^2 + (2.33)^2} = 370. \end{aligned}$$

Este segundo caso provee el tamaño máximo de la muestra, igual a 370 hogares.

2. Una federación de transportistas permite que ciertos gastos de sus afiliados (gasolina, lubricantes y lavado) se hagan mediante la utilización de la tarjeta de crédito expedida para el pago en las gasolineras locales. La empresa ha expedido 10 050 tarjetas.

Para realizar una investigación sobre la utilización de la tarjeta. Se realizó una encuesta preliminar de 90 tarjetas y se encontró que 63 de ellas fueron utilizadas para pagar servicios en el mes en referencia. Se determinó que el total de gastos cancelados con las tarjetas fue de 23 900 dólares y la desviación estándar de 60. Se desea determinar el tamaño de la muestra, con un error del 2 % y una confianza del 95.5 % para estimar: a) la proporción de afiliados que utilizan la tarjeta; b) el gasto promedio mensual cancelado con la tarjeta; c) De los tamaños muestrales anteriores, ¿cuál se aconsejaría?

Solución: Se tiene que $N = 10\,050$ y $n = 90$.

a) Calculemos todos los elementos previos: $\widehat{p} = \frac{63}{90} = 0.7$, $\widehat{q} = 0.3$, $E_p = 0.02$. De aquí,

$$\begin{aligned} n &= \frac{(z_{\alpha/2})^2 N\widehat{p}\widehat{q}}{NE_p^2 + (z_{\alpha/2})^2\widehat{p}\widehat{q}} \\ &= \frac{4(10\,050)(0.7)(0.3)}{(10\,050)(0.02)^2 + 4(0.7)(0.3)} = 1737. \end{aligned}$$

b) El gasto promedio de las 63 tarjetas es de $\overline{x} = \dfrac{23\,900}{63} = 379.37$.
El error es de $E_\mu = 0.02 \times 379.37 = 7.587$ y $s = 40$. El tamaño de la muestra es de

$$\begin{aligned} n &= \frac{(z_{\alpha/2})^2 N s^2}{N E_\mu^2 + (z_{\alpha/2})^2 s^2} \\ &= \frac{4(10\,050)(60)^2}{10\,050(7.587)^2 + 4(60)^2} = 244.1. \end{aligned}$$

c) El tamaño óptimo de la muestra es de 1737 tarjetas, puesto que es el mayor valor calculado.

13.3. Ejercicios

Estimación del total y de la media poblacionales

1. Se seleccionó una muestra aleatoria de $n = 65$ de una población de 400 individuos. La media y la varianza resultaron $\overline{x} = 84.2$ y $s^2 = 170.8$. Para un nivel de confianza del 95.5 %,

    a) calcule los intervalos para el total y para la media poblacionales;

    b) Encuentre el tamaño de muestra necesario para tener un error máximo de 2 cuando se estime la media.

2. Se quiere estimar cuántas balas se gastaron en una práctica de combate en la que intervinieron 8000 efectivos. Para tal efecto, se tomó una muestra de los registros del número de balas empleadas por 115 militares participantes, resultando un promedio de 94.7. Además, en prácticas similares, se ha medido una desviación estándar de 12.1. Para un nivel del 97 %:

    a) Encuentre el intervalo de confianza para el total de balas empleadas en la práctica;

    b) Determine el tamaño de la muestra para que el error máximo sea de 15 000 balas.

3. Una empresa de alquiler de fotocopiadoras desea conocer el total de copias que sus clientes realizan en un mes. De sus registros, que indican que la empresa tiene alquiladas 280 máquinas, se seleccionó 33. En éstas encontró que en un mes se realizaron un promedio de 1228 copias, con una desviación estándar de 193.

    a) Calcule un intervalo de confianza para el total de copias de nivel 95 %;

    b) ¿A cuántas máquinas habrá que incluir en la muestra para tener un error de 21 000 copias?

4. En un estudio sobre el gasto medio mensual en medicinas y salud, en una ciudad de 25 000 familias se encontró, a través de una encuesta a 201 familias, que ellas tenían un gasto promedio de 84.4 dólares, con una desviación estándar de 5.6 dólares. Para un nivel de confianza del 98 %:

    a) encuentre el intervalo de confianza para el gasto familiar en salud, e interprételo;

    b) calcule el tamaño de la muestra para tener un error menor a 1.3 dólares en la estimación.

5. El dueño de un restaurante desea conocer el consumo medio de los clientes de su negocio. De entre las 1469 facturas correspondientes a las ventas que tuvo la última semana seleccionó, al azar, a 119. En ellas determinó un gasto promedio de 4.5 dólares y una desviación estándar de 0.93 dólares.

    a) Halle el intervalo de confianza al 94.5 % para el gasto medio de los clientes;

b) Determine el tamaño de la muestra para que el error sea menor a 270 dólares en la estimación del total;

c) Realice el intervalo de confianza, de nivel 99 %, para el consumo total semanal en el restaurante.

6. En un estudio de mercado se tomó una muestra de 34 personas de clase media, quienes indicaron que gastaban 48 dólares mensuales en diversiones, con desviación estándar de 1.76 dólares. Con una confiabilidad del 98.5 %, halle:

   a) el tamaño muestral mínimo para realizar el mismo estudio en un grupo similar que cuenta con 5000 personas, si el límite para la estimación del error es igual a 0.5 dólares;

   b) el tamaño muestral mínimo para el total del gasto por el grupo de estudio, si el límite para el error de estimación es de 1500 dólares;

   c) el intervalo de confianza para la media del gasto, si los valores del gasto promedio y de la desviación estándar se mantienen en 48 y 1.76 dólares, respectivamente, y se encuesta a 67 personas.

7. El gerente de recursos humanos quiere estimar el número medio y el total de horas anuales de entrenamiento para los 280 empleados de una división de la compañía. Toma información de los registros del año anterior de 35 empleados y obtiene un promedio de 125 y una desviación estándar de 20 horas de capacitación anual.

   a) Calcule los intervalos de confianza, al 99.7 %, para la media y el total de horas empleadas en entrenamiento;

   b) Con los datos anteriores y si la estimación de la media ha de estar a ± 4.5 horas del valor verdadero, o la estimación del total ha de estar a ± 1700 horas del valor verdadero, ¿cuál es el tamaño muestral requerido?

Estimación de la proporción poblacional

8. Se seleccionó una muestra aleatoria de 121 observaciones, siendo cada una un éxito o un fracaso, de una población de 440 elementos.

   a) Si el número de éxitos fue de 40, calcule un intervalo de confianza al 95.5 % para la proporción de éxitos;

   b) Halle el tamaño de la muestra para que el error máximo sea del 6 %, empleando la proporción muestral hallada previamente;

   c) Halle el tamaño máximo de la muestra, si no se tiene información preliminar de la proporción.

9. En un estudio sobre medios de comunicación se quiere estimar la proporción de los estudiantes de un colegio secundario que miran regularmente las noticias. Entre los 3100 estudiantes del colegio se escogieron a 250 para que respondan la pregunta. De éstos, 76 indicaron que miran las noticias. Para un nivel de confianza del 98 %:

   a) encuentre el intervalo de confianza para p;

   b) fije el tamaño de la muestra necesaria para tener un error menor al 5 % si, (i.) se toma el valor de $\widehat{p}$ estimado antes. (ii.) si no se tiene idea previa de la proporción.

10. El Sr. Vargas está pensando postular a la alcaldía de Pelileo. Antes de formalizar su candidatura decide realizar una encuesta de opinión en la localidad. En el cantón hay 12 000 potenciales electores y se realizó una primera consulta a 270 personas, detectándose un apoyo del 30 % de los posibles electores.

a) Encuentre un intervalo de confianza al 97 % para la proporción de votantes que apoyarían al Sr. Vargas e interprete el resultado.

Como el Sr. Vargas no estaba seguro de los resultados de la primera consulta, decide contratar a una empresa para que realice un segundo estudio. La empresa indica que la encuesta tiene un costo fijo de 5000 dólares más un costo variable de 4 dólares por cada entrevista, ¿cuánto le costará este trabajo al Sr. Vargas si él quiere tener un error de 4 % con una confiabilidad del 98 %,

b) si se toma como porcentaje de votantes favorables a la candidatura al 30 %?;

c) si no se tiene una idea previa de la popularidad del Sr. Vargas?

11. Una empresa de asesoría política ha sido contratada para determinar la popularidad de un candidato a alcalde de una ciudad de 215 000 habitantes adultos. En un sondeo realizado con 215 posibles votantes registró un nivel de aceptación del 34 % para el político.

    a) Halle un intervalo de confianza al 96.5 % para la proporción de simpatizantes del candidato;
    b) Si se quisiera realizar otra encuesta utilizando la información disponible, encuentre el tamaño máximo de la muestra a utilizar para que el error sea de 0.03;
    c) Si se conoce que el candidato, históricamente, ha tenido una aceptación de alrededor del 40 %, encuentre el tamaño de la muestra a emplearse.

12. Para efectos de planeación económica en la provincia de Cotopaxi es necesario realizar un estudio entre 2200 hatos ganaderos. Una encuesta piloto arrojó las siguientes estimaciones:

    - Promedio de vacas por hato, 46 y desviación estándar, 20.
    - Rendimiento promedio de leche por hato: 345 litros y varianza de 9700.
    - El 60 % de los hatos tiene un rendimiento medio de leche superior a 250 litros.

    Con un error del 8 % y una confianza del 95.5 %, determine los tamaños de las muestras si se va a estimar:

    a) el número de vacas lecheras por hato;
    b) el rendimiento medio de leche por hato;
    c) la proporción de hatos con un rendimiento promedio de leche superior a 250 litros;
    d) ¿Qué tamaño de muestra usted recomendaría?

La selección de la muestra

Los procedimientos expuestos en esta sección se basan en el principio de que las muestras deben constituirse por elementos tomados al azar, de suerte que cada elemento de la población tenga la misma probabilidad de figurar en la muestra. Solo así ésta puede representar a la población, reproducir fielmente los diversos caracteres y quedar sometida a los procedimientos estadísticos descritos.

Entonces, nos planteamos el siguiente problema práctico a solucionar: hallar un medio de asegurarse de que todos los elementos de la muestra sean tomados al azar.

Teóricamente, no habría problema alguno. En una investigación ideal disponemos del marco muestral, así sabemos cuántos y quiénes son tales elementos. Únicamente habría que sortearlos y los favorecidos serían incluidos en la muestra. Hasta antes de la popularización de las computadoras la selección de las muestras se realizaba con el empleo de tablas de números aleatorios. En la actualidad esta tarea

se realiza con la ayuda de programas informáticos, en los cuales hay comandos especiales que generan números aleatorios y facilitan la selección de la muestra.

Sin embargo, no siempre se puede realizar la extracción de la muestra de la manera descrita, ya sea porque la población es bastante grande y la localización de cada elemento elegido es muy laboriosa o porque el marco muestral no está bien definido. Para superar las dificultades se han ideado otros métodos de muestreo, que a continuación los veremos.

13.4. Muestreo aleatorio estratificado

En el diseño de una muestra, mediante el método aleatorio simple, se presentan varios inconvenientes cuando la población es bastante heterogénea o si está distribuida muy ampliamente; por lo tanto, el número de consultas necesarias para obtener información confiable es alto, con el consiguiente aumento en el costo de la investigación y en el tiempo de ejecución.

Un procedimiento adoptado para superar estos problemas es el de formar una *muestra estratificada* y al azar. Según este método, se subdivide la población en varios grupos, llamados *estratos*, cada uno de los cuales debe ser internamente homogéneo. En cada estrato, se escogen al azar las unidades muestrales a investigar, como ya se explicó; es decir, para cada estrato se aplica muestreo aleatorio simple.

Los estratos pueden formarse partiendo de divisiones geográficas (provincias, ciudades, centros urbanos y rurales, etc.) o bien del sexo de las personas, su edad, la profesión, el nivel socioeconómico, entre otros.

La razón del empleo de este método reside en el hecho de que permite obtener, generalmente, resultados más precisos que aquellos que se consiguen con el método aleatorio simple. Sin embargo, es necesario conocer la esencia de cada estrato para proceder, en cada uno de ellos, a la elección al azar de los elementos de la muestra.

Otra ventaja del método es que además de combinar la información de las submuestras de los estratos, para obtener inferencias de la población completa, también permite utilizar la información de las submuestras para conocer el comportamiento de cada estrato en particular, y compararlos entre sí.

Al igual que hicimos en la sección anterior, examinaremos los casos de estimación del total, la media y la proporción poblacionales.

Supongamos que se divide a la población en K estratos, cada uno de los cuales consta de N_i elementos (con $i = 1, 2, \ldots, K$), por lo que $N = N_1 + N_2 + \cdots + N_K$. Dentro de cada estrato elegiremos n_i elementos que constituirán la muestra.

13.4.1. Estimación del total poblacional

El intervalo de confianza al $(1-\alpha) \times 100\,\%$ para el total poblacional τ es

$$\left(\widehat{\tau}_{est} - z_{\alpha/2}\sqrt{\sum_{i=1}^{K} N_i^2 \left(\frac{N_i - n_i}{N_i}\right)\frac{s_i^2}{n_i}}; \widehat{\tau}_{est} + z_{\alpha/2}\sqrt{\sum_{i=1}^{K} N_i^2 \left(\frac{N_i - n_i}{N_i}\right)\frac{s_i^2}{n_i}}\right).$$

Donde

N es el número de elementos en la población.

N_i es el número de elementos en el estrato i, $i = 1, 2, \ldots, K$.

n_i es el número de elementos de la muestra en el estrato i, $i = 1, 2, \ldots, K$.

$\overline{x}_i$ es el promedio de la muestra seleccionada en el estrato i, $i = 1, 2, \ldots, K$.

s_i^2 es la varianza de los datos de la muestra en el estrato i, $i = 1, 2, \ldots, K$.

$\widehat{\tau}$ es la estimación del total poblacional como $\widehat{\tau}_{est} = N\overline{x}_{est}$.

$\overline{x}_{est}$ es el promedio estratificado, que se calcula por

$$\overline{x}_{est} = \frac{1}{N}(N_1\overline{x}_1 + N_2\overline{x}_2 + \cdots + N_K\overline{x}_K) = \frac{1}{N}\sum_{i=1}^{K} N_i\overline{x}_i.$$

Cuando el tamaño muestral de los n_i es pequeño con respecto al tamaño de los estratos N_i, la fórmula se reduce a

$$\left(\widehat{\tau} - z_{\alpha/2}\sqrt{\sum_{i=1}^{K} N_i^2\frac{s_i^2}{n_i}}; \widehat{\tau} + z_{\alpha/2}\sqrt{\sum_{i=1}^{K} N_i^2\frac{s_i^2}{n_i}}\right).$$

Ejemplo. Los directivos de un centro educativo desean conocer el gasto total de los padres de familia en útiles escolares. Para facilitar la investigación se dividió a los alumnos según su nivel, en preprimario, primario y secundario. Una tabla resumen de cómo se estableció la muestra y los datos hallados, se exhibe a continuación.

	NIVEL		
	Preprimario	Primario	Secundario
No. de alumnos (N_i)	1569	832	647
Tamaño muestral (n_i)	167	121	95
Gasto promedio (dólares) ($\overline{x}_i$)	88.4	131.6	97.0
Varianza muestral (s_i^2)	525	468	700

A un nivel del 95.5 %: a) Encontrar el error muestral en la investigación; b) Construir el intervalo de confianza para el gasto total de los padres de familia del plantel.

Solución: Para nuestros datos tenemos $N = 3048$ y $K = 3$.

a) El error de la estimación es

$$\begin{aligned} E_\tau &= z_{\alpha/2}\sqrt{\sum_{i=1}^{K} N_i^2\left(\frac{N_i - n_i}{N_i}\right)\frac{s_i^2}{n_i}} \\ &= 2\sqrt{6\,919\,295 + 2\,287\,008 + 2\,626\,324} = 6880. \end{aligned}$$

b) Calculemos el promedio estratificado:

$$\begin{aligned} \overline{x}_{est} &= \frac{1}{N}(N_1\overline{x}_1 + N_2\overline{x}_2 + N_3\overline{x}_3) \\ &= \frac{1}{3048}[(1569)(88.4) + (832)(131.6) + (647)(97)] = 102.01. \end{aligned}$$

El total se estima por

$$\widehat{\tau}_{est} = N\overline{x}_{est} = (3048)(102.01) = 310\,950.$$

Con todo lo anterior, el intervalo buscado es

$$\begin{aligned}(\widehat{\tau}_{est} - E_\tau; \widehat{\tau}_{est} + E_\tau) &= (310\,950 - 6880; 310\,950 + 6880)\\ &= (304\,070; 317\,830).\end{aligned}$$

Entonces, se estima que el gasto total en útiles escolares, de todos los padres de familia está entre los 304 070 y 317 830 dólares.

13.4.2. Estimación de la media poblacional

El intervalo de confianza al $(1-\alpha)\times 100\,\%$ para la media poblacional μ es

$$\left(\overline{x}_{est} - \frac{z_{\alpha/2}}{N}\sqrt{\sum_{i=1}^{K} N_i^2\left(\frac{N_i - n_i}{N_i}\right)\frac{s_i^2}{n_i}}; \overline{x}_{est} + \frac{z_{\alpha/2}}{N}\sqrt{\sum_{i=1}^{K} N_i^2\left(\frac{N_i - n_i}{N_i}\right)\frac{s_i^2}{n_i}}\right).$$

Cuando el tamaño muestral de los n_i es pequeño con respecto al tamaño de los estratos N_i, la fórmula se reduce a

$$\left(\overline{x}_{est} - \frac{z_{\alpha/2}}{N}\sqrt{\sum_{i=1}^{K} N_i^2\frac{s_i^2}{n_i}}; \overline{x}_{est} + \frac{z_{\alpha/2}}{N}\sqrt{\sum_{i=1}^{K} N_i^2\frac{s_i^2}{n_i}}\right).$$

Ejemplo. Una empresa que realiza estudios de la audiencia televisiva desea estimar el tiempo medio de horas diarias que pasan frente al televisor las familias, en un cantón específico. La empresa dividió la zona de estudio en urbana y rural y se escogió una muestra del 2 por mil; es decir, 2 de cada mil familias de cada área pertenecerán a la muestra. Con los datos resumidos en la siguiente tabla, encuentre un intervalo de confianza al 97.5 % para el tiempo medio que cada familia pasa frente a la televisión.

Area	No. de familias N_i	Tamaño de la muestra n_i	Tiempo promedio $\overline{x}_i$	Varianza muestral s_i^2
Urbana	64 796	130	2.93	0.77
Rural	42 188	84	1.46	0.58

Solución: Calculemos el promedio estratificado:

$$\begin{aligned}\overline{x}_{est} &= \frac{1}{N}(N_1\overline{x}_1 + N_2\overline{x}_2)\\ &= \frac{1}{106\,984}[(64\,796)(2.93) + (42\,188)(1.46)] = 2.35.\end{aligned}$$

Como el tamaño de las submuestras, en cada estrato, es pequeño con relación al respectivo total de elementos, emplearemos la fórmula aproximada

$$\begin{aligned}E_\mu &= \frac{z_{\alpha/2}}{N}\sqrt{\sum_{i=1}^{K} N_i^2\frac{s_i^2}{n_i}}\\ &= \frac{2.24}{106\,984}\sqrt{(64\,796)^2\left(\frac{0.77}{130}\right) + (42\,188)^2\left(\frac{0.58}{84}\right)} = 0.1276.\end{aligned}$$

Por tanto,

$$\begin{aligned}(\overline{x}_{est} - E_\mu; \overline{x}_{est} + E_\mu) &= (2.35 - 0.1276; 2.35 + 0.1276) \\ &= (2.22; 2.48).\end{aligned}$$

Entonces, el tiempo medio que cada familia gasta frente a la televisión está entre las 2h 13 min y 2h 29 min, aproximadamente.

13.4.3. Estimación de la proporción poblacional

El intervalo de confianza al $(1-\alpha) \times 100\,\%$ para la proporción poblacional p es

$$\left(\widehat{p}_{est} - \frac{z_{\alpha/2}}{N}\sqrt{\sum_{i=1}^{K} N_i^2 \left(\frac{N_i - n_i}{N_i}\right)\frac{\widehat{p}_i\,\widehat{q}_i}{n_i - 1}}; \widehat{p}_{est} + \frac{z_{\alpha/2}}{N}\sqrt{\sum_{i=1}^{K} N_i^2 \left(\frac{N_i - n_i}{N_i}\right)\frac{\widehat{p}_i\,\widehat{q}_i}{n_i - 1}}\right)$$

Donde

N es el número de elementos en la población.

N_i es el número de elementos en el estrato i, $i = 1, 2, \ldots, K$.

n_i es el número de elementos de la muestra en el estrato i, $i = 1, 2, \ldots, K$.

$\widehat{p}_i$ es la proporción muestral del estrato i, $i = 1, 2, \ldots, K$.

$\widehat{p}_{est}$ es la proporción estratificada, que se calcula por

$$\widehat{p}_{est} = \frac{1}{N}(N_1\widehat{p}_1 + N_2\widehat{p}_2 + \cdots + N_K\widehat{p}_K) = \frac{1}{N}\sum_{i=1}^{K} N_i\widehat{p}_i.$$

Cuando el tamaño muestral de los n_i es pequeño con respecto al tamaño de los estratos N_i, la fórmula se reduce a

$$\left(\widehat{p}_{est} - \frac{z_{\alpha/2}}{N}\sqrt{\sum_{i=1}^{K} N_i^2 \frac{\widehat{p}_i\,\widehat{q}_i}{n_i - 1}}; \widehat{p}_{est} + \frac{z_{\alpha/2}}{N}\sqrt{\sum_{i=1}^{K} N_i^2 \frac{\widehat{p}_i\,\widehat{q}_i}{n_i - 1}}\right).$$

Ejemplo. Refiriéndonos al ejemplo anterior, en tal estudio también se preguntó si las familias miraban regularmente una telenovela dada. Las proporciones muestrales de quienes tienen preferencia por el citado programa se dan en la siguiente tabla:

Area	No. de familias N_i	Tamaño de la muestra n_i	Proporción muestral del estrato $\widehat{p}_i$
Urbana	64 796	130	0.19
Rural	42 188	84	0.24

Estimar la proporción de todas las familias en el cantón que miran la telenovela. Use $1 - \alpha = 0.025$.

Solución: En primer lugar calculemos la proporción estratificada:

$$\widehat{p}_{est} = \frac{1}{106\,984}[(64\,796)(0.19) + (42\,188)(0.24)] = 0.2097.$$

Como los tamaños muestrales son pequeños, con respecto a los estratos, $\left(\frac{N_i - n_i}{N_i}\right) \approx 1$, para $i = 1, 2$, utilizaremos la fórmula aproximada

$$\begin{aligned} E_p &= \frac{z_{\alpha/2}}{N}\sqrt{\sum_{i=1}^{2} N_i^2 \frac{\widehat{p}_i\widehat{q}_i}{n_i - 1}} \\ &= \frac{2.24}{106\,984}\sqrt{(64\,796)^2\frac{(0.19)(0.81)}{129} + (42\,188)^2\frac{(0.24)(0.76)}{83}} = 0.0625. \end{aligned}$$

El intervalo de confianza es:

$$\begin{aligned} (\widehat{p}_{est} - E_p; \widehat{p}_{est} + E_p) &= (0.2097 - 0.0625; 0.2097 + 0.0625) \\ &= (0.1472; 0.2722). \end{aligned}$$

Así, entre 14.72 % y el 27.22 % de las familias miran la telenovela.

13.4.4. Tamaño y asignación de la muestra

Puesto que en el método de estratificación se forman varios grupos, se acostumbra a fijar de antemano el nivel del error, de acuerdo a éste se establece el número de elementos que constituirán la muestra y luego se determina cuantos de ellos se asignarán a cada estrato.

Sean N_1, N_2, ..., N_K los elementos incluidos en cada estrato, su suma es igual a N, el total de la población. Se seleccionan n_1, n_2, ..., n_K elementos, tomados al azar, de cada estrato. Tendremos que el total de unidades de la muestra es $n = n_1 + n_2 + \cdots + n_K$. Nuestro problema será determinar el tamaño de la muestra y el número de unidades que se consultarán en cada estrato.

Para la asignación de la muestra se utiliza uno de los siguientes 3 métodos: *equitativa, proporcional* y *no proporcional (de Neyman)*.

Asignación equitativa

En la forma de asignación equitativa en cada uno de los estratos se investiga el mismo número de elementos.

El tamaño de la muestra a ser investigada se calcula por

$$n = \frac{K(z_{\alpha/2})^2 \sum_{i=1}^{K} N_i^2 s_i^2}{E^2 N^2 + (z_{\alpha/2})^2 \sum_{i=1}^{K} N_i s_i^2}.$$

Una vez fijado el tamaño total de la muestra se seleccionan, al azar, el mismo número de unidades muestrales en cada estrato. Este número de unidades se calcula por:

$$n_i = \frac{n}{K}, \quad i = 1, 2, \ldots, K.$$

Observación. En este y en los siguientes casos se debe tener presente que en el caso de la estimación de la proporción la varianza se calcula mediante $s_i^2 = \widehat{p}_i\,(1 - \widehat{p}_i)$.

Asignación proporcional al tamaño del estrato

Si en una investigación se han generado submuestras de igual tamaño, podría suceder que los estratos pequeños estén sobrerepresentados, mientras aquellos con muchos elementos estén subrepresentados.

En la asignación proporcional, la muestra está constituida por un número de elementos, en cada estrato, proporcional al tamaño de éstos, con relación al total; es decir, los estratos mayores serán los que tendrán submuestras de mayor tamaño.

El tamaño de la muestra se calcula por

$$n = \frac{N(z_{\alpha/2})^2 \sum\limits_{i=1}^{K} N_i s_i^2}{E^2 N^2 + (z_{\alpha/2})^2 \sum\limits_{i=1}^{K} N_i s_i^2}.$$

Los tamaños de las submuestras de cada estrato se calculan por:

$$n_1 = N_1 \frac{n}{N}, \quad n_2 = N_2 \frac{n}{N}, \quad \ldots, \quad n_K = N_K \frac{n}{N}.$$

En el muestreo con asignación proporcional el error es mínimo, pero por razones de economía o de facilidad en la recolección de los datos en el sondeo, pudiera ser mejor no hacerlo de esta manera.

Asignación no proporcional al tamaño del estrato (o asignación de Neyman)

Supongamos que la población a examinar está dividida en dos grandes estratos. Si el primero de ellos agrupa a personas muy homogéneas entre sí, con respecto a la característica que se considere, bastará con interrogar a unas pocas personas para deducir, con la precisión suficiente, la magnitud que se investiga. Si, por el contrario, el segundo estrato está formado por personas heterogéneas, será preciso interrogar a un número mayor para obtener resultados que supongan la misma exactitud que en el primer caso.

En la asignación de Neyman, el tamaño de la muestra se calcula por

$$n = \frac{(z_{\alpha/2})^2 \left(\sum\limits_{i=1}^{K} N_i s_i\right)^2}{E^2 N^2 + (z_{\alpha/2})^2 \sum\limits_{i=1}^{K} N_i s_i^2}.$$

Para realizar la asignación de los tamaños de las submuestras se procede de la siguiente manera:

1. Si $s_1, s_2, \ldots, s_K$ son las desviaciones estándar de cada uno de los estratos, se determina el valor T definido por

$$T = N_1 s_1 + N_2 s_2 + \cdots + N_K s_K = \sum_{i=1}^{K} N_i s_i,$$

2. Con ésto, los tamaños muestrales de los estratos se calculan así:

$$n_1 = N_1 s_1 \frac{n}{T}, \quad n_2 = N_2 s_2 \frac{n}{T}, \quad \ldots, \quad n_K = N_K s_K \frac{n}{T}.$$

En el siguiente cuadro se expone un esquema de muestra estratificada según los criterios de asignación equitativa, proporcional y no proporcional.

Estratos	Proporción de los estratos (N_i/N)	Desviación estándar (s_i)	Muestra equitativa	Muestra proporcional	Muestra no proporcional
A	0.40	1	500	800	400
B	0.30	2	500	600	600
C	0.20	3	500	400	600
D	0.10	4	500	200	400
Total	1.00		2000	2000	2000

La muestra con asignación no proporcional está formada por un número de elementos, en cada estrato, que supone dar un mayor peso relativo a los que presentan una mayor variabilidad. Por ello, el estrato A, aunque incluye el 40 % de los elementos de la población, figura en la muestra final con el mismo tamaño que el último (el D), que únicamente comprende el 10 % de la población total. Esto es debido a que la variabilidad del estrato D es el cuádruplo de la que tiene el estrato A.

Una vez establecida la composición de la muestra se procederá al sorteo de los individuos a ser entrevistados, que por lo general se lo hace bajo el criterio de la no reposición; es decir, que un elemento de la muestra no puede ser consultado más que una sola vez.

Ejemplo. Para los datos del ejemplo sobre el tiempo promedio diario que gasta la población viendo la televisión, que a continuación se resume:

Area	**No. de familias**	**Tiempo promedio**	**Varianza muestral**
Urbana	64 796	2.93	0.77
Rural	42 188	1.46	0.58

Considerando un error de 0.1 h, a un nivel de confianza del 95.5 %, determinar los tamaños muestrales y de las submuestras mediante asignación: a) equitativa; b) proporcional; b) no proporcional.

Solución: Se tiene que $N = 106\,984$, $E = 0.1$ y $z_{0.0225} = 2$.

a) **Equitativa**. El tamaño de la muestra es

$$\begin{aligned} n &= \frac{K(z_{\alpha/2})^2 \sum_{i=1}^{K} N_i^2 s_i^2}{E^2 N^2 + (z_{\alpha/2})^2 \sum_{i=1}^{K} N_i s_i^2} \\ &= \frac{2 \times 2^2 \times \left((64\,796)^2 \times 0.77 + (42\,188)^2 \times 0.58\right)}{(0.1)^2 \times (106\,984)^2 + 2^2 \times (64\,796 \times 0.77 + 42\,188 \times 0.58)} = 297.34. \end{aligned}$$

La muestra total es de 298 familias. En cada área se consultará a $\frac{298}{2} = 149$ hogares.

b) **Proporcional**. El tamaño de la muestra es

$$\begin{aligned} n &= \frac{N(z_{\alpha/2})^2 \sum_{i=1}^{K} N_i s_i^2}{E^2N^2 + (z_{\alpha/2})^2 \sum_{i=1}^{K} N_i s_i^2} \\ &= \frac{106\,984 \times 2^2 \times (64\,796 \times 0.77 + 42\,188 \times 0.58)}{(0.1)^2 \times (106\,984)^2 + 2^2 \times (64\,796 \times 0.77 + 42\,188 \times 0.58)} = 277.31. \end{aligned}$$

El número de familias a consultarse es de 278.

Por la fórmula de la asignación proporcional:

$$\begin{aligned} n_1 &= N_1 \frac{n}{N} = 64\,796 \times \frac{278}{106\,984} = 168.34, \\ n_2 &= N_2 \frac{n}{N} = 42\,188 \times \frac{278}{106\,984} = 109.63. \end{aligned}$$

Entonces, se deberá tomar una muestra de 169 hogares en la zona urbana y de 110 en la rural.

c) **De Neyman**. El tamaño de la muestra es

$$\begin{aligned} n &= \frac{(z_{\alpha/2})^2 \left(\sum_{i=1}^{K} N_i s_i\right)^2}{E^2N^2 + (z_{\alpha/2})^2 \sum_{i=1}^{K} N_i s_i^2} \\ &= \frac{2^2 \times \left(64\,796 \times \sqrt{0.77} + 42\,188 \times \sqrt{0.58}\right)^2}{(0.1)^2 \times (106\,984)^2 + 2^2 \times (64\,796 \times 0.77 + 42\,188 \times 0.58)} = 276. \end{aligned}$$

Calculemos el denominador T:

$$T = N_1 s_1 + N_2 s_2 = 64\,796 \times \sqrt{0.77} + 42\,188 \times \sqrt{0.58} = 88\,988.$$

Ahora, apliquemos las fórmulas correspondientes:

$$\begin{aligned} n_1 &= N_1 s_1 \frac{n}{T} = 64\,796 \times \sqrt{0.77} \times \frac{276}{88\,988} = 176.4, \\ n_2 &= N_2 s_2 \frac{n}{T} = 42\,188 \times \sqrt{0.58} \times \frac{276}{88\,988} = 99.7. \end{aligned}$$

La muestra deberá estar formada por 177 familias de la zona urbana y 100 de la zona rural. ◀

En esta sección solo se presentaron las formas de fijar los tamaños muestrales basados en información estadística. Existen otras formas de realizar tal determinación, tomando en cuenta los costos de efectuar la investigación, los costos fijos o el costo unitario de cada toma, éstos no los expondremos.

Determinación de los estratos

El estratificado es el método más utilizado por las empresas y entidades que se ocupan de realizar sondeos, con adaptaciones prácticas que tienen en cuenta los costos y las posibilidades reales de la investigación.

Antes de la confección de la muestra conviene tener en cuenta, ante todo, los fines de la investigación y las características de la población que interesen de modo particular, y que pueden tener una importancia fundamental en las conclusiones que se darán.

Las clasificaciones que más comúnmente se consideran en la elaboración de las muestras son:

1. El sexo de las personas.
2. Los grupos de edad.
3. La región y la dispersión geográfica.
4. El urbanismo o ruralismo de la localidad.
5. El nivel educativo.
6. El nivel socioeconómico, entre otros.

Para fines prácticos, el verdadero y adecuado sorteo, en cada estrato, es frecuentemente impracticable o implica gastos excesivos y pérdida de tiempo. Por ello se recurre al *método de las cuotas*, que consiste en asignar un cierto número de entrevistas que se deben realizar en cada estrato (cuotas), por cada entrevistador.

Compete, entonces, al encargado de la entrevista elegir –al azar– las personas que han de ser interrogadas dentro del ámbito de cada cuota que le ha sido asignada.

Para disminuir, las distorsiones que causaría la falta de aleatoriedad en este método, las empresas suelen aumentar la fiabilidad de sus estudios mediante el «sobremuestreo», o sea la realización de más encuestas que el número originalmente planificado, así se compensaría el aumento del error antes introducido.

13.5. Ejercicios

Estimación del total y de la media poblacionales

1. Utilice los datos de la siguiente tabla para:

   a) hallar un intervalo de confianza al 95.5 % para el total poblacional τ;

   b) hallar un intervalo de confianza al 95.5 % para la media poblacional μ.

Estrato	No. de unidades muestrales N_i	Tamaño de la muestra n_i	Promedio muestral $\overline{x}_i$	Varianza muestral s_i^2
I	2000	200	521	2410
II	3000	200	402	2938
III	1000	200	325	2047

2. Para establecer un sistema de subsidios en el consumo de la energía eléctrica se hizo una investigación por muestreo en una ciudad. Se dividió a los hogares según su nivel socioeconómico (NSE) y se tomó una muestra del 2 %. Los resultados se resumen a continuación.

NSE	No. total de hogares	No. de hogares muestreados	Consumo promedio	Varianza muestral
I	12 425	30	125	232
II	34 871	70	97	175
III	69 724	140	48	124

   Para un nivel de confianza del 96.5 %, halle:

a) una estimación por intervalo para el consumo medio de los hogares;

b) una estimación con un intervalo para el consumo de todos los hogares de la ciudad.

Suponga que se desea realizar otro muestreo en el que se tendrá un error en la estimación de la media de 3 dólares. Determine los tamaños muestrales en cada estrato si la asignación se realiza mediante:

c) asignación equitativa;

d) asignación proporcional;

e) asignación no proporcional.

3. En un sondeo para determinar el gasto anual de la población de una ciudad en arreglo personal, se clasificó a los consultados según su sexo. Los datos se muestran a continuación:

	SEXO	
	Masculino	Femenino
Tamaño del estrato	2500	2300
Tamaño muestral	250	150
Gasto promedio	70	140
Varianza muestral	25	169

a) Determine el gasto promedio y el gasto total, mediante un intervalo de confianza de 93 %;
Encuentre el tamaño de cada estrato, para tener un error de 1.5 y para que la muestra sea realizada mediante:

b) asignación equitativa;

c) asignación proporcional;

d) asignación no proporcional.

4. Se realizó una encuesta para estimar el total de ventas semanales de los locales de productos naturistas de Quito. De los 1415 negocios de este tipo, se escogieron al azar 135. A continuación se resumen los datos recogidos, según la ubicación geográfica de los locales.

Situación	No. de locales	Locales en la muestra	Venta promedio	Varianza
Norte	600	45	478	204
Centro	265	45	413	358
Sur	550	45	394	513

a) Encuentre un intervalo de confianza al 99.7 % para el total de ventas de dichos locales;
Si se quiere realizar un estudio en el cual el error de estimación sea de 5000 dólares, encuentre los tamaños muestrales, en cada estrato, para que la muestra sea realizada mediante:

b) afijación equitativa;

c) afijación proporcional;

d) afijación de Neyman.

5. En una universidad se decidió llevar a cabo un estudio sobre el ahorro que mantienen sus empleados para cuando ellos se retiren. Se tomó una muestra aleatoria estratificada del 10 % de la población, por grupos de edad, con afijación proporcional. Luego de procesar la información, se obtuvieron los siguientes resultados:

Edad	**menos de 40 años**	**40 a 55 años**	**más de 55 años**
No. de empleados	280	150	220
Media	800	1400	3200
Desv. estándar	160	400	750

a) Estime el ahorro medio de los empleados, mediante un intervalo, y obtenga el error de estimación para un nivel de confiabilidad del 96 %;

b) Estime el total ahorrado por los empleados de la universidad.

Si se desearía realizar otro estudio, a un nivel del 94 %, en el cual se quiere tener un error de 100 dólares en la estimación de la media. Encuentre los tamaños muestrales en cada estrato mediante:

c) afijación equitativa;

d) afijación de Neyman.

6. Un comerciante planeó comprar los remates de productos que realizó la aduana. Para obtener un valor aproximado del lote, el comerciante seleccionó aleatoriamente 100 artículos de cada tipo de producto puesto a remate. En la siguiente tabla aparecen el número de artículos según su tipo, el costo promedio de las muestras y sus desviaciones estándar.

Artículo	N_i	$\overline{x}_i$	s_i
Calzado	450	800	200
Ropa	380	560	150
Juguetes	230	940	220

a) Encuentre un intervalo de confianza al 95.5 % para el valor promedio y para el valor total de la compra;

b) El comerciante tiene un capital de 750 mil dólares para realizar la compra, ¿de acuerdo con el resultado anterior, puede decirse que él se decida a comprar el lote?;

El comerciante decide que seleccionará una muestra utilizando la información anterior y se ha fijado un error de 15 000 dólares en la estimación del total. Encuentre los tamaños de cada estrato para que la muestra sea con:

c) asignación equitativa;

d) asignación proporcional;

e) asignación no proporcional.

Estimación de la proporción poblacional

7. Utilice los datos de la siguiente tabla para hallar un intervalo de confianza al 98 % para la proporción poblacional.

	Estratos		
	I	II	III
Tamaño del estrato	1000	1200	700
Tamaño muestral	100	100	100
Proporción muestral	0.32	0.26	0.29

8. En un sondeo electoral para conocer la aceptación de un candidato a prefecto de una provincia se entrevistó a un grupo de electores, previa clasificación según su zona de residencia. La siguiente tabla da un resumen.

Area	N_i	n_i	$\widehat{p}_i$
Urbana	92 000	250	0.43
Rural	88 000	150	0.51

a) Encuentre una estimación, con un intervalo al 94 %, del porcentaje de votación que obtendría el candidato;

b) Según el resultado, ¿podría esperarse que el candidato gane las elecciones por una mayoría?;

Si se desea que el error de estimación sea del 5.5 %, encuentre la composición de la muestra para que ella sea seleccionada mediante:

c) afijación equitativa;

d) afijación proporcional;

e) afijación no proporcional.

9. Se desea establecer el porcentaje de habitantes, en la provincia del Guayas, que tienen fe en San Biritute[2]. Se dividió la zona de estudio en ciudad y campo y se preguntó si creían o no en tal deidad. A continuación se resumen los resultados.

Zona	Población total	Tamaño de la muestra	Proporción muestral
Ciudad	2 800 000	450	0.29
Campo	650 000	350	0.61

Encuentre un intervalo de confianza al 97.5 % para el porcentaje de la población de la provincia del Guayas que es creyente en San Biritute.

10. En una provincia, se realizó una encuesta, entre los niños en edad escolar, para conocer la asistencia a las escuelas. Para el efecto se seleccionaron 150 niños, 50 en cada estrato, y se obtuvo los siguientes resultados:

Condición de pobreza	Población total	Proporción muestral
Indigentes	12 000	0.45
Pobres	36 000	0.60
No pobres	27 000	0.74

a) Calcule el intervalo de confianza al 98 % para la proporción de niños de la provincia que asisten a la escuela;

Si la estimación debe tener un error del 3.5 %, encuentre los tamaños muestrales para que:

b) la muestra sea mediante asignación equitativa;

c) la muestra sea mediante asignación proporcional;

d) la muestra sea por asignación de Neyman.

11. El Sr. Vargas está pensando postular a la alcaldía del Puyo. Antes de formalizar su candidatura decide realizar una encuesta de opinión en la localidad. Para ello se zonificó el cantón en 3 sectores y se obtuvo los siguientes resultados:

Zona	**Total de habitantes**	**Número de consultados**	**Pobladores a favor**
Norte	15 000	200	40
Centro	5000	100	27
Sur	25 000	200	60

A un nivel de confiabilidad del 95 %:

a) Encuentre la estimación de intervalo entre los cuales se podría considerar que se encuentra la popularidad del precandidato;

[2]San Biritute es un santo de la tradición popular del Guayas a quien se le asigna el poder de hacer llover y mejorar las cosechas.

Si la encuesta tiene un costo fijo de 5000 dólares más un costo variable de 4 dólares por cada entrevista, ¿cuánto le costará este trabajo al Sr. Vargas si se quiere tener un error de 5.5 % y la selección será

b) asignación equitativa;

c) asignación proporcional;

d) asignación no proporcional?

12. En una investigación sobre la producción de manzanas en la Provincia del Tungurahua se desea estimar la proporción de agricultores que se dedican al cultivo de la mencionada fruta y la producción media, en miles de kg, de cada parcela. Se realizó un sondeo en 3 cantones de la provincia, a continuación se presenta un resumen de los datos obtenidos.

Cantón	Tamaño de la población	Tamaño de la muestra	Proporción muestral	Promedio	Varianza muestral
Ambato	5135	75	0.65	6 000	640 000
Patate	2173	50	0.71	8 000	722 500
Pelileo	3412	50	0.52	5 000	518 400

a) Encuentre el intervalo de confianza al 96 % para la proporción de agricultores que se dedican al cultivo de la manzana;

b) Encuentre el intervalo de confianza al 96 % para la producción media de manzanas en la zona de estudio;

Si en un estudio se planea realizar 175 encuestas, determine los tamaños muestrales en cada estrato, de manera que la forma de asignación de los estratos sea:

c) equitativa;

d) proporcional al tamaño del estrato;

e) no proporcional al tamaño del estrato.

13.6. Muestreo por conglomerados

La elaboración de un muestreo aleatorio puede ser costoso y difícil de realizar porque la población está dispersa en un área extensa y la localización de cada elemento de la muestra podría llevar mucho tiempo. En estos casos se practica el muestreo por conglomerados.

Definición (de conglomerado) Los conglomerados son subconjuntos de la población que tienen la propiedad de ser internamente lo más heterogéneos y entre ellos lo más homogéneos posible.

Por ejemplo, en una investigación se desea conocer la opinión de las amas de casa de una ciudad. En lugar de sortear a los individuos, se procede a muestrear aleatoriamente las manzanas de la ciudad y después a entrevistar a todas las amas de casa que viven en cada una de las manzanas seleccionadas. Así, cada manzana contendrá un conglomerado de elementos y el número de elementos variará de un conglomerado a otro.

En este tipo de muestreo la construcción del marco muestral es fácil, porque se maneja elementos mayores y los costos de la investigación se rebajan. En cambio, se corre el riesgo de que los elementos en cada conglomerado sean muy homogéneos; por ejemplo, si en una manzana viven únicamente familias de un nivel socioeconómico alto, las respuestas de las amas de casa consultadas pueden ser muy parecidas, perdiéndose la heterogeneidad interna requerida.

Para compensar estos problemas se necesita escoger el número suficiente de conglomerados para tener la necesaria variación en las respuestas.

Al igual que en las otras secciones, examinaremos los intervalos de confianza para el total, la media y la proporción poblacionales.

13.6.1. Estimación del total poblacional

A continuación presentaremos el estimador por intervalo del total poblacional τ y una fórmula aproximada para el cálculo del tamaño de la muestra.

Intervalo de confianza

El intervalo de confianza al $(1-\alpha)\times 100\,\%$ para el total poblacional τ es

$$\left(\widehat{\tau} - z_{\alpha/2} N s\sqrt{\frac{N-n}{Nn}}; \widehat{\tau} + z_{\alpha/2} N s\sqrt{\frac{N-n}{Nn}}\right).$$

Donde

N es el número de conglomerados en la población.

n es el número de conglomerados en la muestra.

m_i es el número de elementos en el conglomerado i, con $i = 1, 2, \ldots, n$.

$\overline{m}$ es el tamaño promedio del conglomerado en la muestra, que se calcula por $\overline{m} = \dfrac{\sum_{i=1}^{n} m_i}{n}$.

M es el número de elementos en la población, que se calcula por $M = \sum_{i=1}^{N} m_i$.

$\overline{M}$ es el tamaño promedio del conglomerado para la población, que se calcula por $\overline{M} = \dfrac{M}{N}$.

x_i es la suma de las observaciones correspondientes al i–ésimo conglomerado, con $i = 1, 2, \ldots, n$.

$\widehat{\tau}$ es la estimación del total poblacional como $\widehat{\tau} = M\overline{x}$.

$\overline{x}$ es el promedio muestral, que se calcula por $\overline{x} = \dfrac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} m_i}$.

s es la desviación estándar, que se calcula por $s = \sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \overline{x}m_i)^2}{n-1}}$

y $\sum_{i=1}^{n}(x_i - \overline{x}m_i)^2 = \sum_{i=1}^{n} x_i^2 - 2\overline{x}\sum_{i=1}^{n} x_i m_i + \overline{x}^2 \sum_{i=1}^{n} m_i^2$.

Nota. Si no se dispone de $\overline{M}$ se utiliza $\overline{m}$.

Tamaño de la muestra

El número de conglomerados a incluir en una muestra, obtenida de una población conformada por N conglomerados, con un 95.5 % de confianza y un error E_τ dado es

$$n = \frac{n_0 N}{n_0 + N}, \quad \text{donde} \quad n_0 = \frac{(z_{\alpha/2})^2 N^2 s^2}{E_\tau^2}.$$

La fórmula de evaluación de n incorpora, en el denominador, una corrección que se debe a que tratamos con una población finita.

Ejemplo. En una ciudad viven 38 300 personas distribuidas en 10 500 familias. Se seleccionaron 12 familias para estimar el gasto mensual en transporte. Los datos se encuentran en la siguiente tabla:

Familia	No. de personas por familia (m_i)	Gasto mensual ($)	Gasto total por familia (x_i)
1	3	25.2; 54.0; 22.0	101.2
2	4	35.0; 18.6; 61.3; 18.0	132.9
3	2	46.2; 45.3	91.5
4	1	53.9	53.9
5	3	71.0; 69.3; 84.0	224.3
6	3	94.0; 78.8; 16.0	188.8
7	2	83.4; 19.3	102.7
8	4	27.7; 94.8; 38.1; 43.0	203.6
9	1	48.6	48.6
10	2	23.3; 58.9	82.2
11	3	73.6; 63.4; 31.0	168.0
12	3	68.5; 65.5; 42.1	176.1

a) Encontrar un intervalo de confianza para el gasto total de la población en transporte, al 98.5 %; b) Manteniendo los datos de a), determinar el tamaño de la muestra para tener un error de 100 mil dólares en la estimación de τ.

Solución: De los datos del enunciado, el número total de conglomerados es $N = 10\,500$ y el número total de habitantes en la ciudad es $M = 38\,300$; además, $z_{0.0075} = 2.43$.

De la tabla se obtiene que $n = 12$ y

$$\sum_{i=1}^{n} m_i = 31, \qquad \sum_{i=1}^{n} x_i = 1573.8.$$

a) Calculemos los componentes del intervalo de confianza para τ:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} m_i} = \frac{1573.8}{31} = 50.767,$$

$$\widehat{\tau} = M\overline{x} = 38\,300 \times 50.767 = 1\,944\,376,$$

$$\sum_{i=1}^{n} x_i^2 = 245\,491.5, \quad \sum_{i=1}^{n} x_i m_i = 4576.5, \quad \sum_{i=1}^{n} m_i^2 = 91.$$

Como

$$\begin{aligned}
\sum_{i=1}^{n}(x_i-\overline{x}m_i)^2 &= \sum_{i=1}^{n}x_i^2-2\overline{x}\sum_{i=1}^{n}x_im_i+\overline{x}^2\sum_{i=1}^{n}m_i^2\\
&= 245\,491.5-2(50.767)(4576.5)+(50.767)^2(91)\\
&= 15\,354.45,\\
s &= \sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x}m_i)^2}{n-1}}=\sqrt{\frac{15\,354.45}{11}}\\
&= 37.36.
\end{aligned}$$

Sustituyendo los valores antes encontrados en la fórmula:

$$\begin{aligned}
E_\tau &= z_{\alpha/2}Ns\sqrt{\frac{N-n}{Nn}}\\
&= 2.43\times 10\,500\times 37.36\times\sqrt{\frac{10\,500-12}{10\,500\times 12}}=275\,020.
\end{aligned}$$

El intervalo queda:

$$\begin{aligned}
(\widehat{\tau}-E_\tau;\widehat{\tau}+E_\tau) &= (1\,944\,376-275\,020;1\,944\,376+275\,020)\\
&= (1\,669\,356;2\,219\,396).
\end{aligned}$$

Así, el gasto mensual total de la población en transporte está entre los 1.67 y 2.22 millones de dólares.

b) Para el cálculo del tamaño muestral usamos $E_\tau=100\,000$. Primero, obtengamos n_0:

$$n_0=\frac{(z_{\alpha/2})^2N^2s^2}{E_\tau^2}=\frac{(2.43)^2(10\,500)^2(37.36)^2}{(100\,000)^2}=90.87.$$

Entonces,

$$\begin{aligned}
n &= \frac{n_0N}{n_0+N}\\
&= \frac{90.87\times 10\,500}{90.87+10\,500}=90.1.
\end{aligned}$$

El tamaño es de 91 conglomerados.

13.6.2. Estimación de la media poblacional

En lo que sigue se presenta el intervalo de estimación de la media poblacional y una forma aproximada de determinar el tamaño muestral.

Intervalo de confianza

El intervalo de confianza al $(1-\alpha)\times 100\,\%$ para la media poblacional μ es

$$\left(\overline{x}-z_{\alpha/2}\frac{s}{\overline{M}}\sqrt{\frac{N-n}{Nn}};\overline{x}+z_{\alpha/2}\frac{s}{\overline{M}}\sqrt{\frac{N-n}{Nn}}\right).$$

Tamaño de la muestra

El número de conglomerados a incluir en una muestra, obtenida de una población conformada por N conglomerados, con un 95.5 % de confianza y un error E_μ dado es

$$n = \frac{n_0 N}{n_0 + N}, \text{ donde } n_0 = \frac{(z_{\alpha/2})^2 s^2}{(\overline{M})^2 E_\mu^2}.$$

Ejemplo. Si en el ejemplo anterior, sobre el gasto en transporte de las familias de una ciudad, se deseara una estimación del gasto mensual promedio en transporte por persona. a) Encontrar un intervalo de confianza para la media poblacional, al 98.5 %; b) Manteniendo los mismos datos, fijar el tamaño de la muestra para tener un error de 5 dólares en la estimación de μ.

Solución:

a) Habíamos determinado que $N = 10\,500$, $M = 38\,300$ y $n = 12$. Además,

$$\sum_{i=1}^{12} m_i = 31, \qquad \sum_{i=1}^{12} x_i = 1573.8,$$
$$\overline{m} = 2.583, \qquad \overline{M} = 3.647,$$
$$\overline{x} = 50.767, \qquad s = 37.36.$$

El error queda:

$$\begin{aligned} E_\mu &= \frac{z_{\alpha/2}\, s}{\overline{M}} \sqrt{\frac{N-n}{Nn}} \\ &= \frac{2.43 \times 37.36}{3.647} \sqrt{\frac{10\,500 - 12}{10\,500 \times 12}} = 7.182. \end{aligned}$$

Consecuentemente, el intervalo de confianza para μ es

$$\begin{aligned} (\overline{x} - E_\mu; \overline{x} + E_\mu) &= (50.767 - 7.182; 50.767 + 7.182) \\ &= (43.585; 57.949). \end{aligned}$$

Lo que significa que el gasto medio de la población en concepto de transporte está entre los 43.59 y los 57.95 dólares mensuales.

b) El tamaño de la muestra para tener un error $E_\mu = 5$ se calcula así:

$$n_0 = \frac{z_{\alpha/2}\, s^2}{(\overline{M})^2 E_\mu^2} = \frac{(2.43)^2 (37.36)^2}{(3.647)^2 (5)^2} = 24.79.$$

Entonces, el tamaño de la muestra es

$$\begin{aligned} n &= \frac{n_0 N}{n_0 + N} \\ &= \frac{24.79 \times 10\,500}{24.79 + 10\,500} = 24.73. \end{aligned}$$

El tamaño requerido es de 25 familias.

13.6.3. Estimación de la proporción poblacional

Presentamos ahora la forma de establecer el intervalo de confianza y el tamaño de la muestra cuando se estima la proporción poblacional mediante conglomerados.

Intervalo de confianza

El intervalo de confianza al $(1-\alpha)\times 100\,\%$ para la proporción poblacional p es

$$\left(\widehat{p} - z_{\alpha/2}\frac{s_p}{\overline{M}}\sqrt{\frac{N-n}{Nn}}; \widehat{p} + z_{\alpha/2}\frac{s_p}{\overline{M}}\sqrt{\frac{N-n}{Nn}}\right).$$

Donde

N es el número de conglomerados en la población.

n es el número de conglomerados en la muestra.

m_i es el número de elementos en el conglomerado i, con $i = 1, 2, \ldots, n$.

$\overline{m}$ es el tamaño promedio del conglomerado en la muestra, que se calcula por $\overline{m} = \dfrac{\sum_{i=1}^{n} m_i}{n}$.

M es el número de elementos en la población, que se calcula por $M = \sum_{i=1}^{N} m_i$.

$\overline{M}$ es el tamaño promedio del conglomerado para la población, que se calcula por $\overline{M} = \dfrac{M}{N}$.

y_i es el total de éxitos en el i–ésimo conglomerado, con $i = 1, 2, \ldots, n$.

$\widehat{p}$ es la proporción muestral, que se calcula por $\widehat{p} = \dfrac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i}$.

s_p es la desviación estándar, que se calcula por $s_p = \sqrt{\dfrac{\sum_{i=1}^{n}(y_i - \widehat{p}m_i)^2}{n-1}}$

y $\sum_{i=1}^{n}(y_i - \widehat{p}m_i)^2 = \sum_{i=1}^{n} y_i^2 - 2\widehat{p}\sum_{i=1}^{n} y_i m_i + \widehat{p}^2 \sum_{i=1}^{n} m_i^2$.

Nota. Si no se dispone de $\overline{M}$ se utiliza $\overline{m}$.

Tamaño de la muestra

El número de conglomerados a incluir en una muestra, obtenida de una población conformada por N conglomerados, con un $(1-\alpha)\times 100\,\%$ de confianza y un error E_p dado es

$$n = \frac{n_0 N}{n_0 + N}, \text{ donde } n_0 = \frac{z_{\alpha/2}^2\, s_p^2}{(\overline{M})^2 E_p^2}.$$

Ejemplo. Para conocer el uso de Internet por los alumnos de una universidad se seleccionó aleatoriamente a 17 de los 498 cursos de la entidad. A los alumnos se les preguntó si en la última semana habían utilizado los servicios de Internet. A continuación se da la información respecto al número de alumnos consultados en cada curso y el número de respuestas afirmativas.

Curso (i)	No. de consultas (m_i)	No. respuestas afirmativas (y_i)	Curso (i)	No. de consultas (m_i)	No. respuestas afirmativas (y_i)
1	25	13	10	66	45
2	34	16	11	78	51
3	56	20	12	29	19
4	87	33	13	35	20
5	21	11	14	48	22
6	36	26	15	27	16
7	45	23	16	64	39
8	44	18	17	54	48
9	51	27			

a) Hallar el intervalo de confianza al 97 % para la proporción de estudiantes de la universidad que han utilizado Internet; b) ¿A los alumnos de cuántos cursos hay que consultar para tener un error del 4 % en la estimación?

Solución: De los datos de la tabla se obtiene que $N = 498$, $n = 17$ y

$$\sum_{i=1}^{n} m_i = 800, \qquad \sum_{i=1}^{n} y_i = 447.$$

Por tanto, la estimación de la proporción p es

$$\widehat{p} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i} = \frac{447}{800} = 0.5588.$$

a) Calculemos los otros elementos del intervalo:

$$\sum_{i=1}^{n} y_i^2 = 14245, \qquad \sum_{i=1}^{n} y_i m_i = 24\,006, \qquad \sum_{i=1}^{n} m_i^2 = 43\,336.$$

Entonces,

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \widehat{p}m_i)^2 &= \sum_{i=1}^{n} y_i^2 - 2\widehat{p}\sum_{i=1}^{n} y_i m_i + \widehat{p}^2 \sum_{i=1}^{n} m_i^2 \\
&= 14\,245 - 2(0.559)(24\,006) + (0.559)^2(43\,336) \\
&= 947.86, \\
s_p &= \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{p}m_i)^2}{n-1}} = \sqrt{\frac{947.86}{16}} \\
&= 7.697.
\end{aligned}$$

Para el cálculo del error no se dispone del valor de $\overline{M}$, por lo que emplearemos $\overline{m}$:

$$\begin{aligned}
E_p &= \frac{z_{\alpha/2}\, s}{\overline{M}}\sqrt{\frac{N-n}{Nn}} \\
&= \frac{2.17 \times 7.697}{47.06}\sqrt{\frac{498-17}{498 \times 17}} = 0.0846.
\end{aligned}$$

Entonces, el intervalo es

$$\begin{aligned}(\widehat{p}-E_p;\widehat{p}+E_p) &= (0.5588-0.0846;0.5588+0.0846)\\ &= (0.4742;0.6434).\end{aligned}$$

Es decir que entre el 47.4 % y el 64.3 % de los estudiantes han accedido a Internet la última semana.

b) En el cálculo del tamaño empleamos $E_p = 4\,\% = 0.04$; así,

$$\begin{aligned}n_0 &= \frac{(z_{\alpha/2})^2 s_p^2}{(\overline{M})^2 E^2}\\ &= \frac{(2.17)^2(7.697)^2}{(47.06)^2(0.04)^2} = 78.73.\end{aligned}$$

El tamaño de la muestra es

$$\begin{aligned}n &= \frac{n_0 N}{n_0+N}\\ &= \frac{78.73\times 498}{78.73+498} = 67.98.\end{aligned}$$

Entonces, se debe consultar en 68 cursos de la universidad.

13.7. Ejercicios

Estimación del total y de la media poblacionales

1. Como resultado de un muestreo por conglomerados se obtuvo la siguiente información:

$$\begin{aligned}N &= 10\,000, \qquad n=100,\\ m_1 &= m_2=\cdots=m_{100}=15, \qquad M=150\,000,\\ \sum_{i=1}^{n} x_i &= 4800, \qquad \sum_{i=1}^{n}(x_i-\overline{x}m_i)^2 = 474.\end{aligned}$$

a) Encuentre un intervalo de confianza al 95.5 % para el total poblacional τ;

b) Encuentre un intervalo de confianza al 95.5 % para la media μ.

2. En una ciudad pequeña, donde hay 3200 hogares, se realizó un sondeo para estimar el tiempo medio que en los hogares se ve la televisión diariamente. La empresa dividió la ciudad en 450 manzanas, porque no disponía de una lista de los hogares y seleccionó 8 manzanas de casas al azar. En la siguiente tabla se indican el número de hogares y el tiempo total (en minutos) que las familias, en cada manzana, dedican a mirar la televisión.

Manzana (i)	No. de hogares (m_i)	Tiempo total (x_i)	Manzana (i)	No. de hogares (m_i)	Tiempo total (x_i)
1	5	1120	5	8	2610
2	6	1350	6	7	2780
3	8	2720	7	7	2550
4	12	3560	8	9	2770

a) Halle los intervalos de confianza al 94 % para el tiempo total y el tiempo medio que los hogares de la ciudad dedican a ver la televisión;

b) Encuentre el tamaño de la muestra para tener un error de 19 minutos en la estimación de la media.

3. En una academia dedicada a la enseñanza del inglés se desea estudiar el número de años que los alumnos llevan aprendiendo este idioma. Se realizó esta pregunta en 5 de las 68 clases que tiene la academia con los resultados siguientes:

Clase 1: 5, 4, 7, 5, 6, 5.
Clase 2: 8, 6, 8, 9, 6, 7, 10, 8, 6.
Clase 3: 8, 12,10, 11, 12, 9, 13, 12.
Clase 4: 12, 9, 11, 10, 10, 8, 12, 10, 13, 12
Clase 5: 14, 12, 10, 11, 12, 13, 13.

a) Estime el número medio de los años que llevan estudiando inglés los estudiantes de la academia. Use $\alpha = 4\,\%$;

b) Si se quisiera estimar el promedio de años que llevan estudiando todos los estudiantes que se encuentran en academias de enseñanza de inglés que funcionan en la ciudad, con un error de 0.5 años y confiabilidad del 97 %, utilizando como muestra piloto la consulta realizada a los 5 cursos anteriores, encuentre el tamaño de la muestra, si se estima que todas las academias tienen abiertos un total de 725 clases.

Estimación de la proporción poblacional

4. Como resultado de un muestreo por conglomerados se ha obtenido la siguiente información:

$$\begin{aligned} N &= 10\,000, \qquad n = 100, \\ m_1 &= m_2 = \cdots = m_{100} = 15, \qquad M = 150\,000, \\ \sum_{i=1}^{n} y_i &= 300, \qquad \sum_{i=1}^{n}(y_i - \widehat{p}m_i)^2 = 5628. \end{aligned}$$

Encuentre un intervalo de confianza al 95.5 % para la proporción poblacional p.

5. Un importador de ropa usada, después de recibir un cargamento de 500 paquetes, seleccionó aleatoriamente 10 de ellos y contó el número de prendas defectuosas por paquete. Basados en los datos de la siguiente tabla:

Paquete (i)	No. de prendas (m_i)	No. de prendas defectuosas (y_i)	Paquete (i)	No. de prendas (m_i)	No. de prendas defectuosas (y_i)
1	85	23	6	66	15
2	44	20	7	78	25
3	56	18	8	59	19
4	77	23	9	55	27
5	61	21	10	68	22

a) Encuentre un intervalo de confianza al 96 % para la proporción p de prendas defectuosas en el cargamento de ropa;

b) Establezca el tamaño mínimo para tener un error de la estimación del 3 %.

6. Una empresa agroexportadora empaca melones en cajas que contienen 8 unidades cada una. Por problemas en el transporte se estropeó un cargamento de 1000 cajas. Para conocer las pérdidas ocasionadas se seleccionó aleatoriamente 20 cajas y se contó el número de melones golpeados en cada caja. Los resultados se dan a continuación:

Caja	No. de frutas golpeadas	Caja	No. de frutas golpeadas
1	0	11	0
2	1	12	1
3	1	13	0
4	2	14	2
5	0	15	0
6	0	16	0
7	2	17	3
8	5	18	1
9	3	19	0
10	0	20	4

a) Encuentre un intervalo de confianza al 96.5 % para la proporción de melones golpeados;

b) Si el exportador pierde 1.7 dólares por cada fruta estropeada, encuentre el intervalo de confianza al 96.5 % para la pérdida total del cargamento;

c) ¿Cuál debe ser el tamaño de la muestra para tener un error de 0.045?

7. En un barrio de la ciudad viven 200 familias. Una muestra de 9 familias suministró información relativa al número de miembros, ingreso familiar quincenal (en dólares), gasto en alimentación (en dólares) y si tiene suscripción a algún periódico. Los resultados fueron:

Familia	No. de miembros	Ingreso quincenal	Gasto en alimentación	Suscripción periódico
1	2	262	82	si
2	3	184	96	no
3	3	193	101	no
4	5	145	76	no
5	4	283	122	si
6	7	301	113	si
7	2	247	104	no
8	4	255	123	no
9	2	221	102	si

Con una confiabilidad del 95 %, estime:

a) el ingreso quincenal medio de los miembros de las familias;

b) el gasto promedio en alimentación por persona;

c) el ingreso quincenal promedio de las familias que tienen suscripción, si se sabe que en el barrio hay 75 suscritos. Compare con los resultados de a);

d) la proporción de familias suscriptoras a un periódico. Compare con el dato real;

e) Con los datos de a), b) y c), determine el tamaño óptimo de la muestra, si se desea tener un error de estimación del 15 %.

Capítulo 14

Respuestas

Capítulo 1

Sección 1.7

3. N: 40 %, R:15 %, B: 45 %.
4. a) categóricos; c) el 21 %.
5. a) tamaño, peso, velocidad y agresividad son datos ordinales; función es nominal.
6. 2.5 %.
8. c) 12.5 %.
9. 26 %.
10. 20 %.
11. 52.4 %.
12. 70 %.
13. $n = 160$.
14. a) 880; b) 752.
15. $n = 50$.
16. 76.
17. 275.
18. 44.64 %.
19. 18 %.
20.

Intervalo	0 – 40	40 – 80	80 – 120	120 – 160	160 – 200
Frec. relativa	0.05	0.45	0.10	0.10	0.30

Sección 1.12

1. La mediana.
2. a) $\overline{x} = 4.875$, b) $Me = 5$; c) $s = 1.553$; d) $R = 5$; e) $As = -0.644$; f) $Ap = 0.592$.
3. a) $\overline{x} = 6$; b) $Q_2 = 5$; c) $s = 2$; d) $R = 5$; e) $RIQ = 4$; f) $As = 0.362$; g) $Ap = -1.826$.
4. c) $\overline{x}_d = 0.75$; $Med_d = 0.35$; $s_d = 1.789$; $\overline{x}_l = 2.33$; $Med_l = 1.75$; $s_l = 2.002$.
5. 2.94 %.
6. a) 6; c) 86 %; d) 66 %.
7. a) **Cuantitativos**: precio, proporción de malta y tiempo de añejamiento, **Cualitativos**: categoría y nota; c) Precio: $\overline{x} = 87.56$, $Mo = 70$, $Q_2 = 86$, Malta: $\overline{x} = 48.96$, $Mo = 100$, $Q_2 = 40$. Tiempo: $\overline{x} = 9.06$, $Mo = 12$, $Q_2 = 8.5$; d) Precio: $s = 23.166$, $RIQ = 28$, $CV = 0.2646$, Malta: $s = 30.114$, $RIQ = 42.5$, $CV = 0.615$, Tiempo: $s = 2.697$, $RIQ = 5$, $CV = 0.298$; e) Precio: $As = 1.339$, $Ap = 2.801$, Malta: $As = 1.092$, $Ap = -0.542$, Tiempo: $As = -0.233$, $Ap = -1.406$.
8. $\overline{x} = 32.54$; $Me = 32$, $Mo = 34$.
9. a) $\overline{x} = 6.56$; b) $Mo = 7$; c) $s = 2.727$; d) $R = 8$.
10. a) $Q_2 = 6$; b) $\overline{x} = 5.046$; c) $s = 2.63$; d) $RIQ = 3$.
11. b) Sugerencia: Use $L_0 = 24.5$ y $A = 4$.
12. a) $\overline{x} = 167.5$, $s = 7.265$; b) $\overline{MH} = 167.189$, $\overline{MG} = 167.344$; c) $Q_2 = 167.5$, $RIQ = 5$.
13. a) $Q_1 = 2.9$, $Q_2 = 5.4$, $Q_3 = 8.05$; b) $\overline{MH} = 3.715$; c) $\overline{MG} = 4.5837$.
14. a) $\overline{x} = 12.29$, $Q_1 = 7.73$, $Q_2 = 11.91$, $Q_3 = 17.36$.
15. mín $= 0$, $Q_1 = 2$, $Q_2 = 3$, $Q_3 = 4$, máx $= 8$.
16. a) $\overline{x}_c = 21.571$, $s_c = 2.149$; b) $\overline{x}_f = 70.829$, $s_f = 3.8685$; c) $\overline{x}_f = 1.8\overline{x}_c + 32$, $s_f = 1.8s_c$.
17. a) $s = 1.348$, $s \approx 1.15$; b) (2.28; 5.52), (1.48; 6.87), (0.13; 8.22); c) $n_1 = 32$, $n_2 = 49$, $n_3 = 50$.
18. b) $Q_2 = 116$; c) $Q_1 = 107.25$, $Q_3 = 121.5$; d) $\overline{x} = 114.59$, $s = 12.43$; f) $As = -0.602$, $Ap = 1.213$.
19. 33.3 %.
20. 665.6 km.
21. 85.62 km/día.
22. 168 cm.
23. 8.
24. 101.
25. 19.
26. $Med = 37.92$, $Mo = 36.92$.
27. 500, la mediana y moda.
28. 20.

29. a) $n = 9$; b) 8.1 cigarrillos.

30. 268 800 Tm.

31. a) 356 mil dólares; b) 49.5.

32. 80 %.

33. $5n$.

34. a) 850; b) 900.71.

35. a) 640; b) 2 %.

36. 6.27.

37. $\overline{x} = 6$, $Q_2 = 6.375$, $s = 2.94$.

38. $\overline{x} = 47$, $Q_2 = 48.33$, $s = 13.75$.

39. $Q_2 = 52.23$.

40. a) 20; b) 30 %; c) $\bar{x} = 1108$, $Med = 1083.3$, $Mo = 1040$.

Capítulo 2

Sección 2.6

1. a) 6; b) 1; c) 3; d) 3.
2. a) 12; b) 1; c) 6; d) 30.
3. a) 3! = 6; b) 4! = 24; c) 5! = 120; d) 6! = 720.
4. a) 12; b) 20; c) 40; d) 65.
5. a) 40, b) 700.
6. a) 9; b) 25; c) 49; d) 64.
7. a) Combinaciones: 10, variaciones: 60; b) Combinaciones: 20, variaciones: 120.
8. a) Son 20 parejas sin reposición y 25 parejas con reposición; b) Son 30 parejas sin reposición y 36 parejas con reposición.
9. De 36 maneras.
10. 1330.
11. 120.
12. 720.
13. 504.
14. 84.
15. 0.5.
16. a) 120; b) 5040; c) 2160.

17. a) $\mathbf{V}_7^2 = 42$; b) $\mathbf{C}_7^2 = 21$.

18. 380.

19. 128.

20. 450 formas diferentes.

21. 105.

22. 84.

23. a) $\mathbf{C}_{16}^6$; b) 16^6; c) 15×16^5; d) 15^3.

24. 17.

25. $26 \sum_{i=0}^{7} 36^i = 2.096 \times 10^{12}$.

26. a) $4 \times 7^5 = 67\,228$; b) $7^3 + 3 \times 7^4 + 4 \times 7^5 = 74\,774$.

27. 6720.

28. a) $x \in A^c \cap B^c \cap C^c$; b) $x \in A \cup B \cup C$; c) $x \in (A^c \cap B^c) \cup (B^c \cap C^c) \cup (A^c \cap C^c)$;
d) $x \in (A^c \cap B \cap C) \cup (A \cap B^c \cap C) \cup (A \cap B \cap C^c)$; e) $x \in (A \cap B \cap C)^c$.

29. Sugerencia: Utilice los axiomas de la probabilidad.

30. $\Pr(A \cap B) = \frac{1}{6}$, $\Pr(A \cup B) = \frac{23}{36}$, $\Pr(A \cap B^c) = \frac{1}{3}$.

31. a) $\frac{1}{2}$; b) $\frac{1}{6}$; c) $\frac{3}{8}$.

32. a) $\Pr(A^c) = 0.625$, $\Pr(B^c) = 0.5$; b) $\Pr(A \cup B) = 0.75$; c) $\Pr(A^c \cap B^c) = 0.25$;
d) $\Pr(A^c \cap B) = 0.375$, $\Pr(A \cap B^c) = 0.25$.

33. Sugerencia: Utilice la fórmula de la probabilidad de la unión.

34. $\Omega = \{1C, 1E, 2C, 2E, 3C, 3E, 4C, 4E, 5C, 5E, 6C, 6E\}$.

35. a) $\Pr(A) = 0.4$; b) $\Pr(A \cup B) = 0.9$; c) $\Pr(B^c) = 0.4$; d) $\Pr(A \cap B) = 0.1$; e) $\Pr(A \setminus B) = 0.3$;
f) $\Pr(A^c \cap B^c) = 0.1$; g) $\Pr[(A \cap B)^c] = 0.9$; h) $\Pr(A \cup B^c) = 0.5$.

36. a) $\Omega = \{(I,I,I), (I,I,D), (I,D,I), (D,I,I), (I,D,D), (D,I,D), (D,D,I), (D,D,D)\}$;
b) $C = A \setminus B$; c) $B^c = \{(I,I,D), (I,D,I), (D,I,I), (I,D,D), (D,I,D), (D,D,I)\}$;
$B \cup C = \{(I,I,I), (I,I,D), (I,D,I), (D,I,I), (D,D,D)\}$; $A \cap B = \{(I,I,I)\}$;
$A^c \cap B^c = \{(I,D,D), (D,I,D), (D,D,I)\}$.

37. a) $\Pr(A) = \frac{1}{3}$; b) $\Pr(A \cup B) = \frac{2}{3}$; c) $\Pr(A^c) = \frac{2}{3}$.

38. $\frac{29}{40}$.

39. a) $\frac{1}{3}$; b) $\frac{1}{2}$.

40. a) $6\left(\frac{1}{4}\right)^3 = \frac{3}{32}$; b) $\frac{9}{64}$; c) $\frac{37}{64}$; d) $\frac{36}{64}$.

41. $\dfrac{1}{15}$.

42. $\dfrac{\mathbf{C}_4^3}{\mathbf{C}_6^3} = \dfrac{1}{5}$.

43. $\dfrac{3}{38}$.

44. $\dfrac{\mathbf{C}_5^0\,\mathbf{C}_{95}^{10} + \mathbf{C}_5^1\,\mathbf{C}_{95}^9}{\mathbf{C}_{100}^{10}} = 0.9231$.

45. $\dfrac{23}{114}$.

46. a) $1 - \dfrac{\mathbf{C}_{96}^3}{\mathbf{C}_{100}^3}$; b) $1 - \dfrac{\mathbf{C}_{99}^3}{\mathbf{C}_{100}^3}$.

47. $1 - \dfrac{\mathbf{C}_{10}^0\,\mathbf{C}_{70}^5}{\mathbf{C}_{80}^5}$.

48. a) $\dfrac{4}{\mathbf{C}_{40}^5}$; b) $\dfrac{30\,\mathbf{C}_{10}^4}{\mathbf{C}_{40}^5}$; c) $\dfrac{40\,\mathbf{C}_{36}^2}{\mathbf{C}_{40}^5}$.

49. $\dfrac{17}{33}$.

50. a) $\dfrac{1}{\mathbf{C}_{44}^6}$; b) $\dfrac{\mathbf{C}_6^4\,\mathbf{C}_{38}^2 + \mathbf{C}_6^5\,\mathbf{C}_{38}^1 + \mathbf{C}_6^6\,\mathbf{C}_{38}^0}{\mathbf{C}_{44}^6}$.

51. a) $\dfrac{567}{31\,250}$; b) $\dfrac{255}{256}$; c) $\dfrac{9}{2 \times 10^7}$; d) $\dfrac{3}{25 \times 10^5}$.

52. $\dfrac{1}{4}$.

53. 0.2913.

54. $\dfrac{20 - 3\pi}{20}$.

55. $p = \dfrac{1}{\pi}$.

Sección 2.9

3. a) $\Pr(A|B) = \dfrac{3}{4}$; b) $\Pr(B|A) = \dfrac{1}{2}$; c) $\Pr(A^c|B) = \dfrac{1}{4}$; d) $\Pr(B^c|A) = \dfrac{1}{2}$; e) $\Pr(A^c|B^c) = \dfrac{5}{8}$; f) $\Pr(B^c|A^c) = \dfrac{5}{6}$.

4. $\dfrac{1}{4}$.

5. a) $\Pr(A) = \dfrac{1}{2}$; b) $\Pr(B) = \dfrac{1}{2}$; c) no son independientes.

6. a) $\Pr(B \setminus A) = \dfrac{7}{18}$, $\Pr(B|A) = \dfrac{1}{2}$, $\Pr(A \cup B^c) = \dfrac{11}{18}$; d) Si son independientes.

7. a) A y B son independientes. b) A y C no son independientes.

8. $\frac{1}{5}$.

9. $\frac{3}{28}$.

10. Envío conjunto: a) 0.9; b) 0.9. Envío por separado: a) 0.81; b) 0.99.

11. $\frac{20}{21}$.

12. 0.75

13. a) $\frac{45}{46}$; b) $\frac{99}{110}$.

14. $\Pr(C|D) = 0.75573$.

15. $\frac{19}{2018}$.

16. 979

17. a) 15.9 %; b) 0.5283.

18. a) 0.08; b) 0.24; c) 0.4; d) 0.8421.

19. a) 0.322; b) 0.258; c) 0.238; d) 0.375.

20. a) 0.24; b) 0.4; c) Si son independientes; d) Si son independientes.

21. a) $\Pr(M) = 0.45$, $\Pr(V) = 0.375$; b) 0.075; c) $\frac{1}{3}$; d) $\frac{2}{3}$; e) no son independientes.

22. a)

Sexo	Hombre	Mujer
p	0.5	0.5

Hábito	No fuma	Fuma	Ex-fumador
p	0.46	0.26	0.28

b) $\Pr(F) = 0.26$, $\Pr(F|M) = 0.32$; c) no son independientes.

23. a) $\frac{8}{15}$; b) $\frac{4}{5}$; c) no son independientes.

24. a) 80 %; b) 0.59; c) 0.8537.

25. 0.3806.

26. $\frac{3}{4}$.

27. a) 24 %; b) 0.667.

28. a) 0.65; b) 0.4.

29. a) 0.4; b) La B: $\Pr(B|H) = 0.5$.

30. a) $\frac{1}{10}$; b) $\frac{7}{10}$; c) $\frac{1}{7}$.

31. 0.8235.

32. $\frac{32}{47}$.

33. a) 0.898; b) 0.39.

34. a) 0.56; b) 0.7273.

35. a) 0.25; b) 0.25, c) independencia; d) 0.45; e) 0.45; f) 0.1.

36. 0.4539.

37. $\text{Card}(A) = \dfrac{mp}{2}$.

Capítulo 3

Sección 3.3

1. a) discreta, $\{0, 1, \ldots, 100 \times 10^9\}$; b) continua, $[0, \infty[$; c) discreta, $\{0, 1, 2, \ldots\}$; d) discreta, $\{0, 1, 2, \ldots\}$; e) continua, $[0, 24 \text{ horas}]$; f) continua, $[0, \infty[$; g) discreta, $\{0, 1, 2, \ldots, 100\}$; h) continua, $]-\infty, \infty[$.

2. Sugerencia: Procure que sus ejemplos estén relacionados con su actividad diaria.

3. a) $\Pr(X \in A) = \dfrac{1}{2}$, $\Pr(X \notin A) = \dfrac{1}{2}$, $\Pr(X \in B) = \dfrac{1}{3}$, $\Pr(X \notin A) = \dfrac{2}{3}$.

4. $$F(x) = \begin{cases} 0, & \text{si } x < -2; \\ 1/4, & \text{si } -2 \le x < 0; \\ 11/12, & \text{si } 0 \le x < \sqrt{3}; \\ 1, & \text{si } x \ge \sqrt{3}. \end{cases}$$

5.

X	5	50	100
p	0.1	0.3	0.6

6.

X	0	1	2
p	68/95	51/190	3/190

7.

X	0	1	2	3	4
p	0.0384	0.6	0.32	0.016	0.0256

$$F(x) = \begin{cases} 0, & \text{si } x < 0; \\ 0.6, & \text{si } 0 \le x < 1; \\ 0.92, & \text{si } 1 \le x < 2; \\ 0.936, & \text{si } 2 \le x < 3; \\ 0.9616, & \text{si } 3 \le x < 4; \\ 1, & \text{si } x \ge 4. \end{cases}$$

8.

X	0	1	2	3
p	0.00001	0.00167	0.07663	0.92169

9. a) $k = \dfrac{3600}{5269}$; b) $\Pr(1 < X \le 4) = \dfrac{1525}{5269}$.

10. a) $c = \dfrac{1}{3}$; b) $\Pr(X < 1) = \dfrac{4}{9}$, $\Pr(X < 2) = \dfrac{2}{3}$, $\Pr(0 < X < 3) = \dfrac{5}{9}$.

11. b) 0.6089.

12. Sugerencia: utilice las propiedades de las series telescópicas.

13. b) $$F(x) = \begin{cases} \sum_{k=0}^{[x]} p(1-p)^k & \text{si } x \ge 0; \\ 0 & \text{si } x < 0. \end{cases}$$

c) $\Pr(X > 2) = (1-p)^3$, $\Pr(X \ge 4) = (1-p)^4$, $\Pr(X < 3) = p(p^2 - 3p + 3)$.

14. a) $c = \frac{3}{38}$; b) $c = 1$; c) $c = \frac{1}{2}$; d) $c = \frac{1}{4}$; e) $c = \frac{1}{\pi}$, f) $c = 1$.

15. a) $\frac{2}{3}$; b) $\frac{1}{3}$; c) $\frac{1}{4}$; d) $\frac{1}{3}$.

16.

X	$-\sqrt{2}$	0	1	$\sqrt{2}$	$\frac{5}{2}$
p	$\frac{1}{8}$	$\frac{11}{40}$	$\frac{1}{10}$	$\frac{1}{4}$	$\frac{1}{4}$

17. a) $F(x) = \begin{cases} 0, & \text{si } x \leq \frac{\pi}{6}; \\ -\cos 3x, & \text{si } \frac{\pi}{6} < x \leq \frac{\pi}{3}; \\ 1, & \text{si } x \geq \frac{\pi}{3}. \end{cases}$

b) $\Pr(X = 0.2) = 0$, $\Pr(X \leq \pi/4) = \frac{\sqrt{2}}{2}$, $\Pr(X > \pi/3) = 0$, $\Pr(\pi/12 \leq X \leq \pi) = 1$.

18. a) $c = 1$, $f(x) = \begin{cases} 1, & \text{si } x \in [0,1]; \\ 0, & \text{si } x \notin [0,1]. \end{cases}$ b) $\Pr(X = 1/3) = 0$, $\Pr\left(X < \frac{1}{3}\right) = \frac{1}{3}$, $\Pr\left(|X| < \frac{1}{4}\right) = \frac{1}{4}$.

19. a) $a = 1$; $b = 2$; b) $F(t) = \begin{cases} 0, & \text{si } t \leq 0; \\ t^3, & \text{si } 0 < t < 1; \\ 1, & \text{si } t \geq 1. \end{cases}$

20. a) $a = \frac{1}{4}$, $b = \frac{1}{2}$; b) $f(x) = \begin{cases} 1/4, & \text{si } x \in [-2,2]; \\ 0, & \text{caso contrario.} \end{cases}$

c) $\Pr(X < 0) = \frac{1}{2}$, $\Pr(|X| < 1.5) = \frac{3}{4}$, $\Pr(|X| > 1.2) = \frac{2}{5}$.

21. a) $\Pr(T > 1) = \frac{1}{2}$; b) $\Pr(T < 2) = \frac{2}{3}$; c) $\Pr(T > 3) = \frac{1}{6}$.

22. a) $F(x) = \begin{cases} 0, & \text{si } x < 0; \\ 0.7, & \text{si } 0 \leq x < 1; \\ 0.9, & \text{si } 1 \leq x < 2; \\ 1, & \text{si } x \geq 2. \end{cases}$ b) $\Pr(X \geq 1) = 0.3$.

23. a)

Y	0	1	2	3	4
Pr	1/25	3/25	5/15	7/25	9/25

b)

G	5	4	3	2	1
Pr	1/25	3/25	5/15	7/25	9/25

24. $F(t) = \begin{cases} 0, & \text{si } t < 0; \\ \frac{t^3(16-3t)}{256}, & \text{si } 0 \leq t \leq 4; \\ 1, & \text{si } t > 4. \end{cases}$ b) $\Pr(T > 2) = \frac{11}{16}$; c) El 26 % de las veces; d) $t = 3.43$.

25. a) $c = \frac{1}{9}$; b) (i.) $\frac{1}{2}$, (ii.) $\frac{3}{4}$; c) $\Pr(A \cap B) = \frac{3}{8}$, A y B son independientes.

26. a) $\Pr(X \leq 3) = 0.316$; b) $\Pr(X \geq 6) = 0.156$, c) $x = 4.81$.

27. a) $\frac{1}{3}$; b) $\frac{11}{18}$; c) no son independientes; d)

Y	1	2	3
Pr	1/9	5/18	11/18

28. a) $F_Y(y) = \begin{cases} 0, & \text{si } y < 0; \\ \sqrt{y}, & \text{si } 0 \leq y \leq 1; \\ 1, & \text{si } y > 1. \end{cases}$

b) $\Pr\left(\frac{1}{16} < X^2 < \frac{1}{9}\right) = \frac{1}{12}$; $\Pr\left(\frac{1}{8} < Y < \frac{5}{6}\right) = 0.559$.

29. $f(t) = \begin{cases} \dfrac{1}{10}, & \text{si } t \in [-5,5]; \\ 0, & \text{si } t \notin [-5,5]. \end{cases}$

30. $\Pr(X^2 < 1) = \dfrac{1}{2}$.

31. $f_U(u) = \begin{cases} \dfrac{1}{3\sqrt{u}}, & \text{si } 0 \le u \le 1; \\ \dfrac{1}{6\sqrt{u}}, & \text{si } 1 < u \le 4; \\ 0, & \text{caso contrario.} \end{cases}$

32. $F_Z(x) = \begin{cases} 0, & \text{si } x < 0; \\ x, & \text{si } x \in [0,1]; \\ 1, & \text{si } x > 1. \end{cases}$ $\quad f_Z(x) = \begin{cases} 0, & \text{si } x \notin [0,1]; \\ 1, & \text{si } x \in [0,1]. \end{cases}$

33. a) $f_Y(x) = \begin{cases} 0, & \text{si } x \le 0; \\ 2\alpha x e^{-\alpha x^2}, & \text{si } x > 0. \end{cases}$ b) $f_Z(x) = \alpha^2 \exp(-\alpha(e^{\alpha x - x}))$ con $-\infty < x < \infty$.

Sección 3.6

1. $\mathbf{E}(X) = 0.161$, $\text{Var}(X) = 0.2153$, $\mathbf{E}(Y) = 4.4$, $\text{Var}(Y) = 2.84$.

2. $\mathbf{E}(X) = 0$, $\text{Var}(X) = 1$, $\sigma(X) = 1$.

3. Si, $\text{Var}(X) > 0$.

4. a) 2515; b) -50; c) 15; d) $2\sqrt{15}$; e) 160; f) 135.

5. a) $\mathbf{E}(S) = 3\mu$, $\text{Var}(S) = 3\sigma^2$; b) $\mathbf{E}(T) = 3\mu$, $\text{Var}(T) = 9\sigma^2$; c) $\mathbf{E}(A) = \mu$, $\text{Var}(A) = \dfrac{\sigma^2}{3}$;

   d) $\mathbf{E}\left(S^2\right) = 3\sigma^2 + 9\mu^2$, $\mathbf{E}\left(A^2\right) = \dfrac{\sigma^2}{3} + \mu^2$.

6. a) $\mathbf{E}(X) = 0$, $\text{Var}(X) = \dfrac{3}{4}$; b) $\mathbf{E}(X) = 0$, $\text{Var}(X) = \dfrac{1}{2}$; c) $\mathbf{E}(X) = 0$, $\text{Var}(X) = \dfrac{1}{6}$;

   d) $\mathbf{E}(X) = 0$, $\text{Var}(X) = \dfrac{3}{5}$.

7. a) $\mathbf{E}(X) = 2.5$, $\sigma(X) = 0.866$; b) $\mathbf{E}(X) = \dfrac{19}{12}$, $\sigma(X) = \dfrac{\sqrt{11}}{12}$; c) $\mathbf{E}(X) = \dfrac{2}{3}$, $\sigma(X) = \dfrac{1}{\sqrt{18}}$;

   d) $\mathbf{E}(X) = -4$, $\sigma(X) = 0.556$; e) $c = \dfrac{1}{\ln 2}$, $\mathbf{E}(X) = \dfrac{1}{\ln 2}$, $\sigma(X) = \sqrt{\dfrac{3}{\ln 4} - \dfrac{1}{(\ln 2)^2}}$.

8. $p = 0.2$, $\alpha = 11$.

9. $\text{Var}(Z) = 0.09$.

10.

X	1	3
p	0.2	0.8

11. $p_1 = 0.35$, $p_2 = 0.20$, $p_3 = 0.45$.

12.

X	1	2	3
p	0.3	0.2	0.5

13. a) $a = 0.5$, $b = 1$; b) $f(x) = 0.5$, si $2 < x \leq 4$; $f(x) = 0$, caso contrario; c) $\Pr(1 < X < 3) = 0.5$, $\Pr(X > 2.5) = 0.25$; d) $\mathbf{E}(X) = 3$, $\text{Var}(X) = \frac{1}{3}$.

14. a) $C = \frac{1}{48}$; b) $\frac{5}{12}$; c) $\Pr(X > 5) = \frac{25}{32}$, $\Pr(X < 7) = \frac{15}{32}$, $\Pr(X^2 - 12X + 35 > 0) = \frac{3}{4}$; d) $\mathbf{E}(X) = \frac{62}{9}$, $\text{Var}(X) = \frac{368}{81}$.

15. a)

X	0	1	2
p	8/15	6/15	1/15

b) $\mathbf{E}(X) = \frac{8}{15}$; c) $\text{Var}(X) = \frac{86}{225}$.

16. $\mathbf{E}(X) = -1.6$ (pérdida).

17. a) $\Omega = \{A, FA, FFA, FFF\}$; b) $\frac{11}{5}$; c) $\mathbf{E}(G) = 1766.4$; d) $\mathbf{E}(X) = \frac{91}{15}$; $\text{Var}(X) = 0.62$

18. $C = 1200$

19. a)

Costo	0	30 000	50 000
p	0.1	0.4	0.5

b) $\mathbf{E}(C) = 37\,000$.

20. a) $\mathbf{E}(X) = 3$ autos. b) $\mathbf{E}(G) = 151$ dólares; c) $\sigma(G) = 90.8$ dólares.

21. $\mathbf{E}(X) = 202.94$; $\sigma(X) = 178.91$

22. a)

X	1	2	3	4	5	6	7	8	9	10
$\Pr(X = i)$	$\frac{19}{100}$	$\frac{17}{100}$	$\frac{15}{100}$	$\frac{13}{100}$	$\frac{11}{100}$	$\frac{9}{100}$	$\frac{7}{100}$	$\frac{5}{100}$	$\frac{3}{100}$	$\frac{1}{100}$

b) $\mathbf{E}(X) = 3.85$; c) $\text{Var}(X) = 5.527$.

23. a) $\mathbf{E}(X) = 109.39$, $\text{Var}(X) = 33.19$; b) 54.70 dólares.

24. a) $\mathbf{E}(X) = 3$, $\sigma(X) = 0.7746$; b) $\mathbf{E}(Y) = 5.5$, $\text{Var}(Y) = \frac{12}{5}$.

25. a) $\mathbf{E}(X) = 2$; $\text{Var}(X) = 0.2$; b) $50.225\,\%$.

26. a) $\mathbf{E}(T) = 4.8\,\text{h}$, $\text{Var}(T) = 2.56\,\text{h}^2$; b) $\mathbf{E}(C) = 48$, $\sigma(C) = 16$.

27. a) $\Pr(X > 60) = 0.334$; b) $\mathbf{E}(X) = 5\pi^2 = 49.348\,\text{m}$; c) $\text{Var}(X) = 461.3\,\text{m}^2$; $\sigma(X) = 21.478\,\text{m}$.

28. Sugerencia: utilice las propiedades de las series geométricas.

29. Sugerencia: utilice el ejercicio anterior.

30. $\mathbf{E}(XY + 2Y - X) = \frac{101}{28}$.

31. a) 2; b) $\frac{7}{3}$; c) $\frac{16}{9}$.

32. $\mathbf{E}(S_n) = 0$, $\text{Var}(S_n) = \frac{n}{2}$.

33. $\mathbf{E}(S_n) = na$, $\mathbf{E}(T_n) = npa$, $\text{Var}(S_n) = n\sigma^2$, $\text{Var}(T_n) = np\left(\sigma^2 + a^2q\right)$.

34. a) $M(t) = \frac{1}{5}\left(e^{-t} + e^{-2t} + e^{-3t} + e^{-4t} + e^{-5t}\right)$, $\mathbf{E}(X) = 3$, $\text{Var}(X) = 2$; b) $M(t) = 0.3e^{-2t} + 0.1e^{-4t} + 0.2e^{-5t} + 0.4e^{-6t}$, $\mathbf{E}(X) = 4.4$, $\text{Var}(X) = 2.84$; c) $M(t) = 0.5e^{-2t} + 0.3e^{2t}$, $\mathbf{E}(X) = 0.4$, $\text{Var}(X) = 3.04$; d) $M(t) = \frac{e^{4t} - e^{-2t}}{6t}$, $\mathbf{E}(X) = 1$, $\text{Var}(X) = 3$; e) $M(t) = \frac{3}{3-t}$, $\mathbf{E}(X) = \frac{1}{3}$, $\text{Var}(X) = \frac{1}{9}$; f) $M(t) = \frac{9e^{\pi t/3} - 3te^{\pi t/6}}{t^2 + 9}$, $\mathbf{E}(X) = \frac{1}{3}(\pi - 1)$, $\text{Var}(X) = \frac{\pi}{9} - \frac{1}{3}$.

Capítulo 4

Sección 4.6

1. a) 15 dólares; b) 240 dólares.

2. 25 dólares.

3. a) $\frac{2}{3}$; b) $\mathbf{E}(X) = 109.5 = 110$ palabras.

4. a) $\mathbf{E}\left(X^k\right) = \frac{(-1)^k + 1}{3}$; b) si k es par, $\mathrm{Var}\left(X^k\right) = \frac{2}{9}$; si k es impar, $\mathrm{Var}\left(X^k\right) = \frac{2}{3}$.

5. a) $\frac{4}{7}$; b) $\frac{20}{7}$; c) $\frac{20}{49}$.

6. $\mathcal{H}(10, 7, 2)$, $\frac{7}{5}$

7. $\Pr(X = 3) = 0.5$, no porque la probabilidad es alta.

8. a) 0.2913; b) 0.8357.

9. a) $\frac{9}{14}$; b) $\frac{13}{14}$; c) 1.

10. a) $\frac{1}{14}$; b) $\frac{13}{14}$; c) $\frac{5}{14}$.

11. a) 0.1536; b) 0.1808; c) 0.9728; d) $\mathbf{E}(X) = 0.8$; e) $\mathrm{Var}(X) = 0.64$.

12. 0.537.

13. a) 0.26214; b) 0.01536.

14. a) 0.36754; b) 0.609.

15. a) 0.0839; b) 0.0374; c) 0.819.

16. a) 0.36; b) 0.997; c) 4.

17. a) 0.9606; b) 0.9994; c) 3.97×10^{-6}.

18. a) 0.05631; b) si es efectiva, pues la probabilidad de que nadie se contagie es muy baja.

19. a) 0.0002187; b) 0.67058; c) 13.

20. a) 0.109; b) 0.9994; c) 0.5885; d) 16; e) 3.2.

21. $\mathbf{E}(G) = 11.6$ millones de dólares. b) $\sigma(G) = 8.08$ millones de dólares.

22. $\Pr(X \leq 95) = 1 - \sum_{k=96}^{100} \mathbf{C}_{100}^{k}(0.05)^k(0.95)^{100-k}$.

23. Al azar: 0.17188; con información: 0.5.

24. 0.0504.

25. a)

X	1	2	3	$\cdots$	k	$\cdots$
p	0.1	0.09	0.081	$\cdots$	$0.1 \times 0.9^{k-1}$	$\cdots$

b) 10 preguntas.

26. a)

X	1	2	3	$\cdots$	k	$\cdots$
p	0.2	0.16	0.128	$\cdots$	$0.2 \times 0.8^{k-1}$	$\cdots$

b) 5 cartuchos.

27. 0.23622.

28. a) 0.2211; b) 0.007946; c) 3 partidos.

29. $\mathbf{E}(X) = 8.33$, es decir 9 refrescos.

30. 3 gusanos.

31. a) 0.04087; b) 0.03564.

32. 0.3188.

33. a) 0.04364; b) 0.00032; c) 25.

34. a) 0.032; b) 5.

35. a) 0.0902; b) 0.9473; c) 0.0527; d) 0.2405.

36. $\lambda = 6$, a) 0.089235; b) 0.061969; c) 0.59278.

37. a) 0.14653; b) 0.56653.

38. a) 0.1954; b) 0.2381; c) 0.7619.

39. a) 0.224; b) 0.5768; c) 0.9502.

40. 0.14288.

41. a) 0.60653; b) Binomial: 0.937; c) 0.96343.

42. 700 dólares.

43. a) 0.3679; b) 0.2652; c) al menos 300.

44. $\mathbf{E}(C) = 200$ dólares, $\sigma(C) = 47.43$ dólares.

45. a) 90.48 %; b) 9.52 %; c) 0.8187.

46. a) 0.0006415; b) $\lambda = 32$, $\Pr(X < 30) = 0.338$.

47. 0.14288.

48. $\lambda = 6$, a) 0.1606; b) 0.1512; c) 0.715.

49. a) $p_0 = 7e^{-2} = 0.94735$; b) 0.12896.

50. a) $\lambda = \frac{26}{20}$; b) 0.4863.

Sección 4.10

1. a) 0; b) 0.75; c) 0.5; d) 0.75; e) $t = -\frac{1}{3}$.

2. a) 0; b) 0.6; c) 0.4; d) 0.8; e) $t = \frac{1}{3}$.

3. $\frac{5}{12}$.

4. 0.6.

5. a) 0.4; b) 0.2; c) $\mathbb{E}(X) = 22\,500$ dólares.

6. a) 0; b) $\frac{7}{12}$; c) $\frac{1}{2}$.

7. a) 0.2; b) $\mathbb{E}(A) = \frac{91}{12}\pi$, $\mathrm{Var}(A) = 6.223$.

8. 0.75.

9. 0.8.

10. 0.15.

11. a) $\lambda = 6$, $f(x) = \begin{cases} 0, & \text{si } x < 0; \\ 6e^{-6x}, & \text{si } x \geq 0. \end{cases}$ $\mathbb{E}(X) = \frac{1}{6}$, $\mathrm{Var}(X) = \frac{1}{36}$;

    b) $\lambda = 3$, $f(x) = \begin{cases} 0, & \text{si } x < 0; \\ 3e^{-3x}, & \text{si } x \geq 0. \end{cases}$ $\mathbb{E}(X) = \frac{1}{3}$, $\mathrm{Var}(X) = \frac{1}{9}$;

    c) $\lambda = 0.5$, $f(x) = \begin{cases} 0, & \text{si } x < 0; \\ 0.5e^{-0.5x}, & \text{si } x \geq 0. \end{cases}$ $\mathbb{E}(X) = 2$, $\mathrm{Var}(X) = 4$;

    d) $\lambda = 0.25$, $f(x) = \begin{cases} 0, & \text{si } x < 0; \\ 0.25e^{-0.25x}, & \text{si } x \geq 0. \end{cases}$ $\mathbb{E}(X) = 4$, $\mathrm{Var}(X) = 16$.

12. a) 0.03; b) 0.66; c) 0.31; d) 0.34.

13. $c = \frac{1}{3}$; a) 0.6321; b) 0.1353; c) 0.2325; d) $\mathbb{E}(T) = 3\,\text{h}$; e) $\mathbb{E}(C) = 120$ centavos.

14. a) 60; b) 0.5034; c) 0.7788.

15. a) 0.368; b) 0.248; c) 0.451; d) 0.452.

16. a) 0.95; b) 0.35.

17. a) 0.287; b) 0.148; c) 0.0695.

18. a) 0.13534; b) 0.2336; c) 16; d) 790.

19. a) 1100; b) 0.135.

20. a) 0.368; b) 0.6065.

21. a) $f(x) = \frac{1}{4\sqrt{2\pi}} e^{-(x-5)^2/32}$; b) $\Pr(Y < 6) = 0.598$, $\Pr(Y > 4) = 0.598$, $\Pr(|Y| < 3) = 0.285$.

22. a) 0.402; b) 0.309; c) 0.276; d) 0.628.

23. 82.28 %.

24. a) 0.788; b) 0.055; c) 0.733.
25. a) 0.02275; b) 0.65542; c) 0.15866.
26. a) 0.861; b) 22 min.
27. 0.403.
28. a) 72; b) 84.6; c) 185.
29. a) 62 cm; b) 58.7 cm.
30. a) 145; b) 537; c) 1484.
31. $h = 2$ metros.
32. a) (i) $\Pr(X > 167) = 0.5$, (ii) $\Pr(X > 170) = 0.16$; b) (i) $\Pr(Y = 4) = (0.16)^4 = 0.0007$, (ii) $\Pr(Y = 2) = \mathbf{C}_4^2(0.5)^4 = 0.375$.
33. a) 0.9088; b) 0.3467; c) 13.49; d) $\Pr(X < 12) = 0.01$, $\sigma = 0.34335$.
34. $\Pr(V) = 0.5987$, $\Pr(M) = 0.6915$, la mujer es relativamente más alta.
35. a) 95.45 %; b) 0.83.
36. a) 0.0062; b) $t = 1172\,\text{h}$.
37. a) 11.5 %; b) 15.8 %; c) 0.00152.
38. a) 0.8413; b) 0.01068; c) 0.6658; d) 0.2367.
39. $\Pr(X > 450) = 0.01$, $\mu = 391.75\,\text{g}$.
40. a) $\mu = 10.72$, $\sigma^2 = 3$.

Sección 4.12

1. 0.023.
2. 0.0985.
3. 0.0003.
4. 192 083.99 kg.
5. 0.0227.
6. a) 0.2375; b) 0.6657.
7. 0.04595.
8. a) 0.97725; b) $n \geq 108$.
9. a) $\mathcal{N}(7.5, 5.25)$; b) 0.1379.
10. 0.0207.
11. a) 0.0823; b) 0.0446.
12. a) 20; b) 0.1269; c) 0.8395.

13. a) 588; b) 0.92758; c) 0.2330.

14. a) 0.2557; b) 0.05593; c) 0.78636.

15. a) $\mathcal{G}(0.36)$, $\mathbf{E}(X) = 2.778$, $\text{Var}(X) = 4.938$;
b) $\mathbf{E}(S_{45}) = 80$, $\text{Var}(S_{45}) = 222.22$;
c) 0.8324.

Capítulo 5

Sección 5.6

1.

Y	X = 0	1	2	3
0	0/30	1/30	2/30	3/30
1	1/30	2/30	3/30	4/30
2	2/30	3/30	4/30	5/30

2. $\frac{1}{20}$; b) $\frac{3}{8}$; c) $\frac{7}{8}$; d) $\frac{23}{120}$;

e)

S	0	1	2
p_S	61/120	51/120	8/120

T	0	1	2	3
p_T	30/120	63/120	21/120	6/120

3. a) $\frac{1}{89}$; b) $\frac{5}{89}$; c) $\frac{29}{89}$; d) $\frac{55}{89}$;

e)

X	-1	0	1	3
p_X	17/89	14/89	17/89	41/89

Y	-1	2	3
p_Y	15/89	27/89	47/89

4. a) $\frac{4}{49}$; b)

X_1	0	1	2
Pr	4/7	2/7	1/7

X_2	0	1	2
Pr	1/7	2/7	4/7

; c) $\frac{40}{49}$.

5.

Y	X = 0	1	2	3	4
$-\sqrt{3}$	0.075	0.050	0.025	0.0375	0.0625
0	0.201	0.134	0.067	0.1005	0.1675
$\sqrt{3}$	0.024	0.016	0.008	0.0120	0.0200

6. a)

X	-3	4	10
p_X	0.25	0.43	0.32

Y	2	4
p_Y	0.55	0.45

b) 0.1965.

7. a)

X	10	20	30	40
p_X	0.14	0.42	0.19	0.25

X	0	1
p_Y	0.29	0.71

b) 0.7392.

8. a) $\frac{1}{60}$; b)

X	0	1	2	3
p_X	1/10	1/5	3/10	2/5

X	1	2	3
p_Y	4/15	1/3	2/5

c) No son independientes;

d) $\Pr(1 \leq X < 3; 2 < Y \leq 3) = \frac{7}{30}$, $\Pr(X + Y < 3) = \frac{1}{10}$.

9. a) $\frac{\sqrt{2}}{4}(\sqrt{3} - 1)$; b) $f_X(x) = \cos x$, $f_Y(y) = \cos y$.

10. $f_X(x) = \frac{\sqrt{10}}{5\sqrt{\pi}} e^{-2x^2/5}$; $f_Y(y) = \frac{\sqrt{2}}{\sqrt{\pi}} e^{-2y^2}$; no son independientes.

11. a) $\frac{4}{\pi^2}$; b) si son independientes.

12. a) $f_X(x) = \lambda^2 x e^{-\lambda x}$, $x > 0$; $f_Y(y) = \lambda e^{-\lambda y}$, $y > 0$; b) $\left(1 - (1+\lambda a)e^{-\lambda a}\right)\left(1 - e^{-\lambda b}\right)$;
c) $1 - (1+\lambda a)e^{-\lambda a}$.

13. a)

X	1	2	3	4	5
p	5/50	13/50	21/50	7/50	4/50

Y	0	1	2	3
p	3/50	22/50	21/50	4/50

b)

$Y\|_{X=2}$	0	1	2
p	1/13	7/13	5/13

c)

$X\|_{Y=0}$	1	2
p	2/3	1/3

14. a)

X	1	2	3	4	5
p	0.22	0.25	0.23	0.20	0.10

Y	1	2	3	4	5	6	7
p	0.30	0.24	0.17	0.11	0.08	0.06	0.04

b)

$X = k$	1	2	3	4	5
$\Pr(X = k\|Y = 1)$	15/30	9/30	4/30	1/30	1/30

15. $\operatorname{Cov}(X,Y) = \dfrac{(n-1)(n+3)(n+1)}{12}$.

16. Si m y n son pares: $\rho(X^m, X^n) = 1$; si m y n son impares: $\rho(X^m, X^n) = 1$; si m y n tiene distinta paridad: $\rho(X^m, X^n) = 0$.

17. a) 0.064; b) 0.102.

18. a) 4; b) $f(x|y_0) = 2x$; c) 0.1089; d) Son independientes.

19. a) $f_X(x) = \begin{cases} \dfrac{2x+4}{5}, & \text{para } 0 \le x \le 1; \\ 0, & \text{caso contrario.} \end{cases}$ b) 0.742; c) 0.04.

20. a) $f(x_1, x_2, x_3) = \begin{cases} \dfrac{(20\,000)^3}{(x_1+100)^3(x_2+100)^3(x_3+100)^3}, & x_1 > 0,\ x_2 > 0,\ x_3 > 0; \\ 0, & \text{caso contrario.} \end{cases}$

b) $\Pr(X_1 < 100, X_2 < 100, X_3 \ge 200) = \dfrac{1}{16}$.

21. a) $k = 1$; b) $f_S(s) = \begin{cases} \dfrac{1}{8}, & \text{si } 0 \le s \le 8; \\ 0, & \text{caso contrario.} \end{cases}$ $f_T(t) = \begin{cases} \dfrac{1}{7}, & \text{si } 0 \le t \le 7; \\ 0, & \text{caso contrario.} \end{cases}$

c) $F(s,t) = \begin{cases} 0 & \text{si } s < 0;\ t < 0; \\ \dfrac{s\,t}{56}, & \text{si } 0 \le s \le 8;\ 0 \le t \le 7; \\ 1, & \text{si } s > 8;\ t > 7. \end{cases}$

22. a) $\dfrac{15}{256}$; b) $f_T(t) = 2t$; c) Son independientes.

23. a) $\dfrac{1}{8}$; b) $f_X(x) = \dfrac{3-x}{4}$, si $0 \le x \le 2$; $f_Y(y) = \dfrac{5-y}{4}$, si $2 \le y \le 4$; c) $\operatorname{Cov}(X,Y) = -\dfrac{1}{36}$.

24. a) 4; b) $f_X(x) = 2xe^{-x^2}$, $f_Y(y) = 2ye^{-y^2}$; c) $\mathbf{E}(X) = \mathbf{E}(Y) = \dfrac{\sqrt{\pi}}{2}$; d) Son independientes.

25. a) $f_X(x) = \dfrac{12x^2+1}{5}$ si $0 \le x \le 1$; $f_Y(y) = \dfrac{4}{5}y(y^2+2)$ si $0 \le y \le 1$; b) $\operatorname{Cov}(X,Y) = \dfrac{-2}{375}$;
c) $\mathbf{E}(X+Y) = \dfrac{209}{350}$, $\mathbf{E}(X^2+Y^2) = \dfrac{27}{25}$.

26. a) $f_X(t) = \frac{1}{\pi}\left(\operatorname{arcsen} t + t\sqrt{1-t^2} + \frac{\pi}{2}\right)$, $f_Y(t) = \frac{2}{\pi}\left(\operatorname{arcsen} t + t\sqrt{1-t^2}\right)$; b) No independientes.

27. a) 8; b) $f_X(x) = \begin{cases} \dfrac{2x+17}{6(1+x)^4}, & \text{si } x \geq 0; \\ 0, & \text{caso contrario.} \end{cases}$ c) No son independientes.

28. a) $k = (n-1)(n-2)$;

b) $F(x,y) = \begin{cases} 1-(x+1)^{2-n}-(y+1)^{2-n}+(x+y+1)^{2-n}, & \text{si } x \geq 0;\ y \geq 0;\ n \geq 0; \\ 0, & \text{caso contrario.} \end{cases}$

29. a) $f_X(x) = \frac{3}{2}(1-x^2)$, $f_Y(y) = 1$, no son independientes;

b) $f_X(x) = \dfrac{2}{(x+1)^3}$, $f_Y(y) = \dfrac{2}{(y+1)^3}$, no son independientes.

30. $f_X(x) = \begin{cases} 6\left(x-x^2\right), & \text{para } 0 \leq x \leq 1; \\ 0, & \text{caso contrario.} \end{cases}$ $f_Y(y) = \begin{cases} 3y^2, & \text{para } 0 \leq y \leq 1; \\ 0, & \text{caso contrario.} \end{cases}$

31. $\text{Cov}(X,Y) = -0.01$, $\rho = -0.1306$.

32. a) 8; b) $f_X(x) = 4(x-x^3)$, $f_Y(y) = 4y^3$; c) $\mathbf{E}(X) = \frac{8}{15}$, $\mathbf{E}(Y) = \frac{4}{5}$; d) 0.6944; e) No son independientes.

33. a) 10; b) $f_X(x) = 5x^4$, $0 < x < 1$, $f_Y(y) = \frac{10}{3}y(1-y^3)$, $0 < y < 1$, no son independientes;

c) $\text{Cov}(X,Y) = \frac{5}{378}$; d) $f_{Y|X=1/2}(y|X=1/2) = 8y$, $0 < y < \frac{1}{2}$; e) $\frac{1}{3}$.

34. a) $f(x,y) = \frac{1}{2}$, si $x \geq 0$, $y \geq 0$, $x+y \leq 2$; b) $f_X(x) = \frac{2-x}{2}$, $f_Y(y) = \frac{2-y}{2}$; c) $\text{Cov}(X,Y) = -\frac{1}{9}$.

35. a) $f(x,y) = \frac{1}{2}$ si $|x|+|y| < 1$; b) $f(x) = 1-|x|$, si $|x| < 1$; c) $\mathbf{E}(X) = 0$, $\text{Var}(X) = \frac{1}{6}$;

d) $\mathbf{E}(XY) = 0$; e) $\text{Cov}(X,Y) = 0$, $\rho(X,Y) = 0$.

36. a) 0.0153; b) 0.0234.

37. a) 0.025; b) 5.344×10^{-4}.

38. a) 0.007; b) 0.013.

39. 0.02923.

40. $\begin{pmatrix} 6 & -2 & 4 \\ -2 & 6 & -4 \\ 4 & -4 & 6 \end{pmatrix}$

41. $\text{Cov}(S_n, T_n) = np\sigma^2$.

42. $\begin{pmatrix} 1 & 2 & 6 \\ 2 & 6 & 18 \\ 6 & 18 & 60 \end{pmatrix}$.

Capítulo 6

Sección 6.5

1. a) 0.7451, b) 0.9976; c) 0.0645; d) 0.0289.
2. $\mu_{\overline{x}} = 3100\,\text{g}$, $\sigma_{\overline{x}} = 15\,\text{g}$.
3. a) $\mu = 400$, $\sigma_{\overline{x}} = 20$; b) 0.9104.
4. 0.0668.
5. a) 0.9890; b) 0.9257; c) 0.81351.
6. 0.13936 exacto, 0.13929 aproximado.
7. 0.15866.
8. 0.9053.
9. 0.06802.
10. $\mu \in (2.87; 3.79)$.
11. $\mu \in (37.77; 40.23)$.
12. a) 0.0456; b) 0.0228.
13. a) 0.0228.; b) 100.
14. 0.0036.
15. 0.923.
16. 0.8132.
17. a) Estadístico; b) 0.00361; c) Si.
18. a) Parámetro, $p = 0.8$; b) El valor de cada estadístico, calculado a partir de las muestras, estará más cercano a 0.8, a medida que aumenta el tamaño de la muestra; c) 0.9846; d) la probabilidad es mayor.
19. 0.785.
20. a) 0.0918; b) 0.0485; c) 0.4772.
21. a) 0.9270; b) 2826.
22. 0.0228.
23. 0.0918.
24. 0.09513.
25. $p \in (0.333; 0.367)$.
26. $p = 0.3679$, $\Pr(\widehat{p} \geq 1/3) = 0.758$.
27. a) $\Pr(\widehat{p} \geq 0.8) = 0.1376$; b) $\Pr(X \geq 20) = 0.1935$; c) La probabilidad de ganar debe mantenerse constante durante el torneo.

28. a) 3.33; b) 3.57; c) 34.17; d) 26.30.

29. 0.95.

30. a) 0.95; b) 0.09.

31. a) $\Pr(s^2 > 8.1) \approx 0.05$; b) $\Pr(2.66 < s^2 < 9.52) \approx 0.94$.

32. $\Pr(s^2 \leq 16) < 0.01$, no es un valor válido para σ^2.

33. $\Pr(s > 7) \approx 0.95$.

34. a) $\Pr(s^2 > 150) \approx 0.9$; b) $\Pr(s^2 > 362) \approx 0.025$; c) $\mathbf{E}(s^2) = 225$, $\text{Var}(s^2) = 3894$.

35. $\sigma^2 \in (49.75; 164.37)$.

36. a) 1.383; b) 2.681; c) 2.086; d) 1.746.

37. 0.8361.

38. a) 0.1; b) 0.75.

39. 0.99.

40. a) 3.32; b) 4.16; c) 4.47; d) 5.86.

41. 0.025.

42. 0.05.

43. $a = \dfrac{1}{2.94} = 0.34$, $b = 2.84$.

44. a) $F \sim F(1,1)$; b) 0.95.

45. a) 0.5885; b) 0.1759.

46. 0.0062.

47. 0.0104.

48. $T \sim t(17)$, 0.05.

49. $T \sim t(22)$, 0.1.

50. 0.0207.

51. 0.9634.

52. 0.0322.

53. 0.9147.

Capítulo 7

Sección 7.5

2. Mejor – $\widehat{\theta}_3$, $\widehat{\theta}_2$, $\widehat{\theta}_1$, $\widehat{\theta}_4$ – Peor.

3. a) Si; b) $\text{Var}(T_1) = \dfrac{\sigma^2}{2n_2}$, $\text{Var}(T_2) = \dfrac{\sigma^2}{n_2}$, $\text{Var}(T_3) = \dfrac{3\sigma^2}{8n_2}$; c) Mejor T_3.

4. a) $\widehat{\theta}_1$, $\widehat{\theta}_2$ y $\widehat{\theta}_4$ son estimadores insesgados; b) El estimador de menor varianza es $\widehat{\theta}_4$.

5. a) $\widehat{\theta}_1$, $\widehat{\theta}_2$ y $\widehat{\theta}_3$ son estimadores insesgados; b) El estimador de menor varianza es $\widehat{\theta}_3$.

7. a) $\widehat{\lambda} = \overline{x}$; b) $\mathbf{E}(V) = 4\mathbf{E}(X) + \mathbf{E}(X^2) = 4\lambda + \lambda^2 + \lambda = 5\lambda + \lambda^2$; c) $\widehat{V} = \dfrac{1}{n}\sum_{i=1}^{n} x_i^2 + 4\overline{x}$.

8. Es insesgado.

9. b) $\alpha = \dfrac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$.

10. b) $\alpha = \dfrac{n_1}{n_1 + n_2}$.

11. a) $\mathbf{E}(\overline{X}_n) = \theta$, $\text{Var}(\overline{X}_n) = \dfrac{1}{3n}$; b) $T_n = \overline{X}_n$, $\mathbf{ECM}(T_n) = \dfrac{1}{3n}$.

12. a) $\widehat{\theta} = 2\left(1 - \dfrac{1}{\overline{X}}\right)$; b) $\widehat{\theta} = 2\left(\dfrac{n-k}{2n-k}\right)$, donde k es el número de veces que aparece el valor 1.

13. $L(p) = 2^{11}p^{29}(1-p)^{21}$, $\widehat{p} = \dfrac{29}{50}$

14. a) $\mathbf{E}(X) = 1 + \theta$, $\widehat{\theta}_1 = \overline{X} + 1$; b) Es insesgado; c) $\widehat{\theta}_2 = \overline{X}$; d) $\widehat{\theta}_2$; e) 1.

15. a) $\widehat{p} = \dfrac{18}{30}$; b) $\widehat{p} = 1 - e^{-600\widehat{\lambda}}$, $\widehat{\lambda} = -\dfrac{1}{600}\ln\left(\dfrac{18}{30}\right)$.

16. a) $2 - \dfrac{4p}{3}$; b) $\widehat{p} = \dfrac{81}{100}$.

17. $\widehat{a} = 0.10129$.

Sección 7.8

1. a) (3.013; 6.986); b) (36.02; 37.98); c) (18.147; 18.852); d) (−22.564; −21.436).
2. a) $L = 0.7052$; b) $L = 53.572$; c) $L = 1.662$; d) $L = 2.773$.
3. a) (117.98; 122.02); b) (220.26; 223.74); c) (65.73; 66.27); d) (−38.17; −35.83).
4. a) (97.9; 102.1); b) (49.19; 56.81); c) (84.42; 87.58); d) (−39.5; −34.5).
5. (47.14; 52.86)
6. a) (416.60; 503.40); b) 295.
7. a) (106.49; 113.51); b) 3.51.
8. a) 0.9938; b) 72.

9. 57.

10. a) (107.1; 108.9); b) 99.068 %.

11. a) (165.01; 169.39); b) 50.

12. a) 50 dólares; b) 490.

13. 62.

14. 43.

15. 31.

16. a) (26.77; 30.04); c) El intervalo hubiera sido más estrecho; d) El intervalo hubiera sido más estrecho; e) El intervalo hubiera sido más estrecho; f) El intervalo hubiera sido de igual longitud, desplazado hacia la izquierda.v

Sección 7.10

1. a) (2.889; 3.471); b) (45.27; 46.73); c) (−1.33; 6.33); d) (−176.58; −87.42).

2. (0.99; 2.30).

3. a) (50.05; 51.38); b) No.

4. a) (16.46, 17.51); b) (0.0396,0.6124).

5. (409.23; 706.76).

6. (219.34; 240.66).

7. 90 %: (375.24; 485.01); 95 %: (363.40; 496.85); 99 % (337.87; 522.38).

8. a) (24.66; 26.86); b) El lote no se acepta.

9. a) (1.89; 2.68); b) El lote no es aceptado.

10. a) (219.34; 236.66); b) $n = 12$: (221.72; 234.28); $n = 100$: (226.19; 229.81).

11. a) (3.527; 3.709); b) $n = 8$.

12. a) (169; 314); b) Si, el tiempo es mayor que el indicado.

13. a) $\overline{x} = 4.846$, $s = 2.949$; b) (3.234; 6.458); c) El intervalo sería más estrecho; d) El ancho del intervalo se mantendrá y se desplaza a la derecha; e) $p = 8/26$. La muestra posiblemente no proviene de una población normal.

Sección 7.12

1. a) (86.43; 1928.57); b) (111.94; 725.81); c) (128.39; 516.25); d) (141.84; 422.43).

2. 90 %: (25.35; 109.17); 95 %: (22.76; 130.61).

3. $\sigma \in (6.61; 17.55)$. La información no es correcta.

4. $\sigma^2 \in (2903.54; 13473.29)$; $\sigma \in (53.88; 116.07)$.

5. $\sigma \in (3.30; 8.40)$. si le conviene.

6. a) (19.245; 23.155); b) (1.51; 34.70).
7. Tradicional: (11.16; 89.70), Alternativo: (9.99; 94.60). Mejor es el método tradicional.
8. $\mu \in (90.115;\ 92.385)$; $\sigma^2 \in (5.43;\ 16.19)$.
9. a) $\mu \in (20.04; 21.13)$; $\sigma^2 \in (0.69; 2.76)$; b) No se recomienda el cambio de variedad.
10. a) $\mu \in (9.23; 20.10)$; b) $\sigma^2 \in (40.89; 176.08)$; c) No se introducirá la nueva línea.

Sección 7.14

1. a) (0.69; 0.81); b) (0.7024; 0.7976); c) (0.6711; 0.8289); d) (0.7108; 0.7892).
2. a) Luis: 0.015, Juan 0.031, Paco, 0.062; b) Diana 0.4, Oliva, 0.2; c) 90 %: C, 95 %: A, 99 %: B; d) La mujer.
3. (0.5341; 0.6659).
4. 2177.
5. a) (0.688; 0.821), no se recomienda su uso; b) 311.
6. a) (0.6556; 0.8043); b) 155.
7. a) (0.2475; 0.3525); b) 276.
8. a) Buenas: (0.865; 0.934), malas: (0.065, 0.134); b) No se aceptará; c) 4238.
9. (0.6587; 0.7319), no se puede introducir la nueva variedad.
10. $E = 0.0278 < 3\,\%$; si son estadísticamente iguales.
11. 97.84 %.
12. a) (0.812; 0.928); b) Será más estrecho; c) Será más ancho.
13. a) $p = 0.5$, $n = 77$; b) Se necesitan más estudiantes; c) Se necesitan más estudiantes.

Sección 7.16

1. $(1.35; 2.65)$, si tiene razón.
2. (4.033; 15.967). La afirmación es cierta.
3. $(-0.68; 6.68)$, son los mismos.
4. $(20 - 61.08; 81.08)$. En cualquiera de las dos ciudades.
5. $(-3.93; 2.10)$, son iguales.
6. (0.93; 7.44), varianzas iguales.
7. a) (0.635; 3.267), son iguales; b) (3.86; 10.14), es mayor el de las mujeres.
8. $(-0.0211;\ -0.1956)$. No son iguales.
9. $(-0.0339;\ 0.0539)$. Si afecta igual.
10. (0.0325; 0.2075). Si es eficaz.

Capítulo 8

Sección 8.6

1. a) H_0: $\mu = 167$, H_1: $\mu > 167$; b) $z_{obs} = 2.366$, se rechaza H_0.
2. $z_{obs} = -1.825$. a) A $\alpha = 0.05$ se rechaza H_0; b) A $\alpha = 0.01$ no se rechaza H_0.
3. a) $z_{obs} = 2.123$, se rechaza H_0, se debe recalibrar la balanza; b) $\alpha = 3.4\,\%$.
4. $z_{obs} = -3$; el costo medio no es 3.
5. $z_{obs} = -4.47$, se rechaza H_0.
6. a) $z_{obs} = -3.16$, se rechaza H_0; b) 0.0008.
7. $z_{obs} = -2.82$. Con $\alpha = 0.01$ se rechaza H_0, ha descendido el número de clientes.
8. a) $z_{obs} = 2.38$; con $\alpha = 0.01$, ha cambiado la demanda; b) $\alpha = 1.72\,\%$.
9. $\beta = 0.1963$.
10. a) $z_{obs} = 4$, la afirmación de la socióloga es falsa; b) **Pot** $= 0 - 8485$.
11. a) $z_{obs} = -3.65$, se rechaza H_0; b) **Pot** $= 0.7088$.
12. a) H_0: $\mu = 18$, H_1: $\mu > 18$; b) Rechazo H_0 si $t_{obs} > t_{0.05}(10) = 1.812$; c) $t_{obs} = 1.976$, rechazo H_0.
13. a) Son estadísticos; b) H_0: $\mu = 26$, H_1: $\mu \neq 26$; c) $t_{obs} = 1.46$, se rechaza H_0 si $t_{obs} > 2.069$; d) La edad de las novias no es diferente de 26 años.
14. $t_{obs} = 3.378$; el periodo medio de espera es diferente a ocho días.
15. $t_{obs} = -1.4785$, no se rechaza H_0.
16. $t_{obs} = 2.56$, se rechaza H_0.
17. $t_{obs} = 1.278$, no se rechaza H_0.
18. $t_{obs} = -2.8247$, se rechaza H_0.
19. a) $t_{obs} = -1.383$; b) $\alpha = 0.1$; c) $\overline{x} = 12.65$ o $s = 0.973$.
20. a) $\mu \in (16.65; 18.36)$; b) $t_{obs} = 2$, la ganancia si es menor que 18.5.
21. a) H_0: $\sigma^2 = 15$, H_1: $\sigma^2 > 15$; b) Se rechaza H_0 si $\chi^2_{obs} > \chi^2_{0.01} = 42.98$; c) $\chi^2_{obs} = 37.44$, no se rechaza H_0. La varianza es igual a 15.
22. $\chi^2_{obs} = 37.44$, no se rechaza H_0.
23. $\chi^2_{obs} = 26.33$. La máquina si garantiza la precisión.
24. $\chi^2_{obs} = 31.97$, se rechaza H_0.
25. a) $t_{obs} = -4.843$, la media de la variable no es 51; b) $\chi^2_{obs} = 16.84$, la varianza es la unidad.
26. a) H_0: $p = 0.3$, H_1: $p < 0.3$; b) Prueba unilateral; c) $z_{obs} = -1.38$, no se rechaza H_0. La proporción es igual a 0.3.

27. a) H_0: $p = 0.2$, H_1: $p > 0.2$; b) $\alpha = 0.04$; c) $z_{obs} = 1.75$, a un nivel de $\alpha = 0.05$ se rechaza H_0.

28. $z_{obs} = -2.0785$, se rechaza H_0.

29. $z_{obs} = 1.417$, el medicamento B no es más efectivo.

30. $z_{obs} = -0.84852$, no se rechaza H_0.

31. a) $z_{obs} = 0.711$, la afirmación del proveedor es cierta; b) $\alpha = 0.2386$.

32. a) $z_{obs} = 1.16$, a un nivel de $\alpha = 0.05$ no se puede decir que la mayoría aprueba el endurecimiento de las penas; b) $\alpha = 0.123$.

33. a) $z_{obs} = 2.26$, se rechaza H_0; b) $\beta = 0.2655$.

34. a) $z_{obs} = -1.854$, la diferencia no es significativa, se debe solo al azar; b) 0.30554.

35. a) $z_{obs} = 3.0$; b) Si es significativo al 1 %; c) Las observaciones provienen de una ley binomial, son obtenidas aleatoriamente y el tamaño es lo suficientemente alto. Las hipótesis si se satisfacen.

Sección 8.10

1. $z_{obs} = 2.635$; se rechaza H_0, las dos máquinas no envasan iguales cantidades.
2. $z_{obs} = -1.434$, no se rechaza H_0.
3. a) $z_{obs} = -2.572$, se rechaza H_0; b) 0.005.
4. $z_{obs} = -0.82$; la media de la Universidad Nacional no es menor que la de la Universidad Técnica.
5. $z_{obs} = 3.677$, se rechaza H_0.
6. $|z_{obs}| = 1.756$; se acepta la igualdad de rendimientos académicos.
7. $z_{obs} = 0.658$, no existe evidencia.
8. $z_{obs} = 2.149$, para $\alpha = 5\,\%$, se rechaza H_0.
9. $|z_{obs}| = 4.205$, se rechaza la igualdad. La norma si permitió el aumento del promedio de goles.
10. $z_{obs} = 1.409$, el consumo es similar.
11. $t_{obs} = 3.735$, si hay aumento en el nivel de plomo.
12. $t_{obs} = 2.588$; se rechaza H_0, si hay evidencia.
13. $t_{obs} = 2.544$. El número de surcos de los mestizos es mayor que el de los aborígenes.
14. a) $|t_{obs}| = 2.115$, el medicamento B es más efectivo; b) $\alpha \approx 0.025$.
15. $t_{obs} = 0.435$. No hay diferencia.
16. $t_{obs} = 2.885$, para $\alpha = 0.01$, se rechaza H_0.
17. $t_{obs} = 1.190$. No se rechaza H_0, la diferencia no es mayor a un minuto.
18. $t_{obs} = 3.939$, la concentración a nivel del suelo es mayor.
19. $|t_{obs}| = 3.254$; se rechaza H_0, si es más lento en la noche.
20. $t_{obs} = 2.236$; se rechaza H_0.

21. $|t_{obs}| = 1.29$, no existe diferencia significativa en los niveles de colesterol.

22. $|t_{obs}| = 1.79$, a nivel 5 %, los dos métodos dan iguales resultados.

23. $t_{obs} = -2.494$, se rechaza H_0.

24. a) $F_{obs} = 1.86$, las varianzas si son iguales; b) Nivel de significación $\alpha = 20\,\%$.

25. a) $F_{obs} = 3.33$, las varianzas no son iguales. La variabilidad puede atribuirse a la máquina 1.

26. a) $F_{obs} = 2.074$, las varianzas si son iguales; b) $t_{obs} = 3.864$, las dos estaciones no pronostican temperaturas iguales.

27. a) $F_{obs} = 2.8$, las varianzas son diferentes; b) $t_{obs} = 4.44$. Si es más efectivo.

28. a) $F_{obs} = 2.945$, se puede asumir varianzas iguales; b) $t_{obs} = 0.5897$, no se rechaza H_0.

29. a) $H_0\colon p_1 - p_2 \leq 0$; $H_1\colon p_1 - p_2 > 0$; b) $z_{obs} = 3.193$; c) Para $\alpha = 0.05$ se rechaza H_0.

30. $|z_{obs}| = 1.275$; no se rechaza H_0, las tasas de desocupación son iguales.

31. a) $|z_{obs}| = 2.041$; no se rechaza H_0, los 2 canales tienen igual nivel de sintonía; b) 0.041.

32. a) $z_{obs} = 1.129$, las dos empresas entregan resultados similares; b) $\alpha = 0.256$.

33. a) $|z_{obs}| = 2.11$, se rechaza H_0; b) $p = 0.0174$.

34. a) Las proporciones muestrales se calculan a partir de dos muestras seleccionadas de manera aleatoria e independiente, de dos poblaciones binomiales, cuyos tamaños son suficientemente altos para que la distribución muestral sea aproximadamente normal; b) $H_0\colon p_1 - p_2 \leq 0.2$; $H_1\colon p_1 - p_2 > 0.2$; $z_{obs} = 0.833$; c) Para $\alpha = 0.05$, el incremento en el hábito de fumar no es mayor que el 20 %.

35. $z_{obs} = 2.619$, se rechaza H_0.

36. $z_{obs} = 1.368$, la diferencia es menor que el 18 %.

Capítulo 9

Sección 9.3

1. $\chi^2_{obs} = 1.654$, no se rechaza H_0.

2. $\chi^2_{obs} = 0.776$, a un nivel de significación del 1 % se acepta H_0. La moneda es simétrica.

3. $\chi^2_{obs} = 7.324$, se rechaza H_0.

4. $\chi^2_{obs} = 3.021$, no se rechaza H_0.

5. a) $\chi^2_{obs}(S) = 19.009$, $\chi^2_{obs}(T) = 16.238$; b) La empresa T dio un mejor resultado.

6. $\chi^2_{obs} = 0.656$, los datos siguen las probabilidades teóricas.

7. a) $\chi^2_{obs} = 12.346$, se acepta H_0. Los números si aparecen uniformemente; b) $\alpha = 0.1945$.

8. $\chi^2_{obs} = 2.699$. El número de hijos varones sigue una ley $\mathbf{Bin}(2, 0.5)$.

9. $\chi^2_{obs} = 4.19$. El número de hijas sigue una ley $\mathbf{Bin}(5, 0.5)$.

10. $\chi^2_{obs} = 4$. Se acepta H_0.

11. $\widehat{\lambda} = 1.027$, $\chi^2_{obs} = 5.506$, no se rechaza que el número de goles sigue una ley $\mathcal{P}(1.027)$.

12. $\widehat{\lambda} = 2.642$, $\chi^2_{obs} = 7.737$, no se rechaza que el número de llamadas sigue una ley $\mathcal{P}(2.642)$.

13. $\chi^2_{obs} = 17.88$. La temperatura no está uniformemente distribuida.

14. $\chi^2_{obs} = 20\,257$; no se acepta H_0.

15. $\widehat{\lambda} = 1/1000$, $\chi^2_{obs} = 14.53$. La duración no sigue la ley $\mathcal{E}(1/1000)$.

16. $\chi^2_{obs} = 4.9$; se acepta H_0.

17. $\chi^2_{obs} = 5.614$; con $\alpha = 0.05$, si hay asociación.

18. $\chi^2_{obs} = 122.33$, no son independientes.

19. a) $\chi^2_{obs} = 1.528$; b) $\chi^2_{obs} = 1.528$; c) No se rechaza H_0.

20. $\chi^2_{obs} = 1.726$; se acepta H_0.

21. a) $\chi^2_{obs} = 30.032$, se rechaza H_0; b) $z_{obs} = 3.146$, se rechaza H_0.

22. $\chi^2_{obs} = 19.12$, se rechaza H_0.

23. $\chi^2_{obs} = 4.876$, se rechaza H_0.

24. b) $\chi^2_{obs} = 1.125$; se acepta H_0.

25. $\chi^2_{obs} = 35.031$; se rechaza H_0.

Sección 9.6

1. a) $p_{obs} = 0.1719$; b) $T_{obs} = 42.5$. No se rechaza H_0.

2. a) $2p_{obs} = 0.3438$; b) $T_{obs} = 20.5$. No se rechaza H_0.

3. $r_{obs} = 12$, no se rechaza H_0.

4. a) $r_{obs} = 9$, no se rechaza la aleatoriedad; b) $2p_{obs} = 0.00028$, se rechaza H_0.

5. a) $r_{obs} = 5$, no se rechaza H_0; b) $2p_{obs} = 0.388$, no se rechaza H_0.

6. $D_{obs} = 0.0268$, no se rechaza H_0.

7. a) $D_{obs} = 0.233$, se acepta H_0; b) $2p_{obs} = 0.754$.

8. a) $r_{obs} = 6$, se acepta H_0; b) $D_{obs} = 0.324$, se acepta H_0;c) $g_{obs} = 1.907$, no hay valores atípicos.

9. a) $r_{obs} = 5$, se acepta H_0; b) $D_{obs} = 0.147$, se acepta H_0; c) $T_{obs} = 33$, no se rechaza H_0; d) $g_{obs} = 1.767$, no hay valores atípicos.

10. a) $r_{obs} = 4$, no rechazo H_0; b) $D_{obs} = 0.1766$, no rechazo H_0; c) $T_{obs} = 0.344$, no rechazo H_0.

11. a) $r_{obs} = 7$, no se rechaza H_0; b) $D_{obs} = 0.287$, no se rechaza H_0; c) $p_{obs} = 0.1938$, no se rechaza H_0.

12. a) $r_{obs} = 6$, no se rechaza H_0; b) $D_{obs} = 0.226$, no se rechaza H_0; c) $T_{obs} = 16$, no se rechaza H_0.

13. $g_{obs} = 17.264$, si es un valor atípico.

14. $x < 141$ o $x > 183$.

15. $2p_{obs} = 0.289$, los dos laboratorios entregan resultados iguales.

16. Prueba de los signos: $2p_{obs} = 0.109$, se acepta H_0, Prueba de Wilcoxon: $T_{obs} = 6$, se rechaza H_0.

17. $U_{obs} = 73$, se acepta la igualdad de niveles de contaminación.

18. $U_{obs} = 11$, se acepta la igualdad de velocidades.

19. $U_{obs} = 41$, los dos tipos de lámparas tienen la misma duración.

20. a) $r_{obs} = 7$, se acepta H_0; b) $D_{obs} = 0.2444$, se acepta H_0; c) $U_{obs} = 40$, se acepta H_0.

21. a) Muestra X: $r_{obs} = 5$, no se rechaza H_0, Muestra Y: $r_{obs} = 6$, no se rechaza H_0;
    b) Muestra X: $2p_{obs} = 1$, $T_{obs} = 4$, no se rechaza H_0, Muestra Y: $2p_{obs} = 0.754$, $T_{obs} = 5$, no se rechaza H_0; c) $U_{obs} = 25.5$, se rechaza H_0, la mediana de X es mayor que la de Y.

22. $r_s = 0.818$, si están correlacionadas.

23. $r_s = 0.429$, no hay asociación.

24. $r_s = 0.864$, se rechaza H_0.

25. a) Prueba de los signos: $2p_{obs} = 0.344$, Prueba de Wilcoxon: $T_{obs} = 9$, se acepta H_0;
    b) $r_s = 0.018$, no están correlacionadas.

Capítulo 10

Sección 10.9

1. $\hat{y} = 2.852 + 3.704x$, $r = 0.948$.

2. a) $\hat{y} = 332.11 + 0.65x$; b) $r = 0.993$; c) $t_{obs} = 16.895$;
   d) $\mathbf{E}(Y) \in (339.63; 344.09)$, $\hat{y}_p \in (335.97; 347.75)$.

3. a) $b_0 = -90.2$, $b_1 = 9$; c)

Fuente	g.l.	SC	MC	F
Regresión	1	810.0	810.0	54.24
Residual	3	44.8	14.933	
Total	4	854.8		

d) $r^2 = 0.948$.

4. b) $y = -7.489 + 0.144x$; c) $F_{obs} = 17.83$; $r^2 = 0.69$; e) No.

5. a) Cantidad =18.976 + 7.271 Tiempo; b)

Fuente	g.l.	S.C.	M.C.	F
Regresión	1	37.01	37.01	79.73
Residual	4	1.857	0.464	
Total corregido	5	38.686		

c) $\beta_0 \in (15.50; 22.45)$, $\beta_1 \in (5.01; 9.53)$.

6. a) $\hat{y} = 3.853 + 0.505x$; b) $s = 3.286$, $\beta_0 \in (-3.55; 11.258)$, $\beta_1 \in (-0.021; 1.031)$; c) $r^2 = 0.64$, $t_{obs} = 2.667$; d) $\hat{y}(12) \in (0.06; 19.77)$; e)

Fuente	g.l.	SC	MC	F
Regresión	1	76.8	76.8	7.11
Residual	4	43.2	10.8	
Total	5	120.0		

7. b) $b = \begin{pmatrix} -4.37 \\ 1.44 \end{pmatrix}$; c) Sí, $F_{obs} = 29.87$; d) $r^2 = 0.888$, $t_{obs} = 14.57$; e) $\mathbf{E}(y_{72}) \in (88.09; 110.43)$; f) $y_{65} \in (53.94; 124.43)$.

8. a) Ganancia $= -225.245 + 2.146$ Préstamos; b) Sí, $r = 0.934$;

c)

Fuente	g.l.	S. C.	C. M.	F
Regresión	1	1735.53	1735.52	41.33
Residual	6	251.97	41.996	
Total	7	1987.50		

d) $\widehat{y}(123) = 38.70$, $y(123) \in (29.54; 47.89)$.

9. a) $y = 70.87 + 35.75x$, $r = 0.735$; b) $y = 11.47 - 1.07x$, $r = -0.848$; c) $y = 459.12 - 34.45x$, $r = -0.896$; d)

Fuente	g.l.	SC	MC	F
Regresión	1	156 965.6	156 965.6	32.63
Residual	8	38 478.9	4809.9	
Total	9	195 444.5		

e) $\beta_0 \in (366.7; 551.5)$, $\beta_1 \in (-487.35; -20.54)$; f) $\mathbf{E}(y) \in (219.09; 320.25)$.

10. b) $b_0 = 4.799$, $b_1 = 0.595$; c) $(44.53; 69.73)$; d) $(30.09; 84.17)$; e) $r = 0.569$, $r^2 = 0.323$.

11. a) $\widehat{y} = 5 + 0.5x$; b)

Fuente	g.l.	S. C.	C. M.	F
Regresión	1	400	400	16.66
Residual	18	432	24	
Total	19	832		

c) $\mathbf{E}(y_{60}) \in (32.36; 37.64)$, $y_{60} \in (24.38; 45.63)$.

12. a) $\widehat{y} = 115.25 - 8.02x$; b) $r^2 = 0.946$; c)

Fuente	g. l.	S. C.	C. M.	F
Regresión	1	229 867	229 867.0	176
Residual	10	13 061	1306.1	
Total	11	242 928		

13. a) En los dos casos se rechaza H_0; b)

Fuente	g.l.	SC	MC	F
Regresión	1	3.210	3.210	73.21
Residual	10	0.439	0.044	
Total	11	3.649		

c) $r^2 = 0.88$.

14. a)

Fuente de variación	g.l.	SC	MC	F
Regresión	1	240	240	12
Residual	7	140	20	
Total	8	380		

b) Se rechaza H_0.

15. $y = e^{-2+3x}$, $r^2 = 1$.

16. a) $\widehat{y} = 0.0687t^{0.0507}$; b) $\widehat{y} = 0.0799 - \dfrac{0.0136}{t}$; c) $r_a^2 = 0.834$, $r_b^2 = 0.876$.

17. a) $\widehat{y} = 3.0348 - \dfrac{7.3306}{x}$; b) $r^2 = 0.9614$.

18. a) $\widehat{y} = 15x^{1.0588}$; b) $\widehat{y} = -12 + 20.8x$; c) $r_a^2 = 0.881$, $r_b^2 = 0.938$.

19. b) $b_0 = 0.5113$, $b_1 = 0.27473$; c) El modelo es adecuado pues $r^2 = 0.916$.

20. b) $y = \exp(4.0657 + 0.1983x)$; c) $r^2 = 0.984$; d) $\widehat{y} = 258$; e) La predicción no es buena.

21. $\ln T = -7.513 + 1.5 \ln r$ o $T^2 = 2.9805 \times 10^{-7} r^3$.

22. $\hat{y} = \dfrac{1}{0.0852 + 0.0375x}$, $r^2 = 0.973$.

23. b) $\hat{y} = -49\,857 + 36\,357x$; c) $r^2 = 0.811$; d) $\hat{y} = 27\,559x$; e) $y = \exp(9.83 + 0.335x)$; f) $r^2 = 0.964$.

Capítulo 11

Sección 11.9

1. a) $y = \dfrac{14}{75} + \dfrac{21}{25}x_1 + \dfrac{4}{75}x_2$; b) $\mathbf{X}^t\mathbf{X} = \begin{pmatrix} 13 & 2 & -2 \\ 2 & 2 & -1 \\ -2 & -1 & 4 \end{pmatrix}$, $\mathbf{X}^t\mathbf{Y} = \begin{pmatrix} 4 \\ 2 \\ -1 \end{pmatrix}$; c) $s^2 = 0.602$;

   d) $R^2 = 0.2396$, $R^2_a = 0.0875$.

2. a) $\mathbf{b} = \begin{pmatrix} 29.124 \\ 24.322 \\ 60.218 \end{pmatrix}$; b) $\hat{y} = 202.8$; c) $R^2 = 0.920$, $R^2_a = 0.866$.

3. a) $\mathbf{b} = \begin{pmatrix} -40.42 \\ 1.37 \\ 4.587 \end{pmatrix}$; b) $R^2 = 0.695$; c)

Fuente	g. l.	*SC*	*MC*	*F*
Regresión	2	24.22	12.1	3.42
Residual	3	10.62	3.5	
Total	5	34.84		

   se acepta H_0;

   d) $t_0 = -1.562$, $t_1 = 1.447$, $t_2 = 1.085$. En los 3 casos se acepta H_0.

4. a) $\mathbf{b} = \begin{pmatrix} -7.690 \\ 0.090 \\ 0.607 \end{pmatrix}$ $\beta_0 \in (-23.47; 8.09)$, $\beta_1 \in (0.032; 0.147)$, $\beta_2 \in (-0.31; 1.245)$; b) $R^2 = 0.961$,

   $R^2_a = 0.921$; c)

Fuente	g. l.	*SC*	*MC*	*F*
Regresión	2	67.65	33.825	24.33
Residual	2	2.78	1.391	
Total	4	70.43		

   d) $\hat{y} = 14.23$, $\mathbf{E}(y_p) \in (11.44; 17.02)$,

   $y_p \in (8.4; 20.02)$.

5. a) $\widehat{G} = 1.083 + 0.01I + 10.749F$; b) $R^2 = 0.856$, $R^2_a = 0.761$, $s^2 = 303.99$; c) $(-0.055;\ 0.074)$;

   d) Solo el tamaño familiar; e) 3.5; f) $F_{obs} = 8.945$, se acepta H_0, a nivel 5 %.

6. a) $\widehat{T} = 32.074 - 0.576L + 0.12A$; b) 16.33; c) $R^2 = 0.948$, $R^2_a = 0.914$;

   d)

Fuente	g. l.	*SC*	*MC*	*F*
Regresión	2	13.521	6.76	27.438
Error	3	0.739	0.246	
Total	5	14.260		

   e) Mejor modelo: $\widehat{T} = 32.337 - 0.529L$.

7. a) $t_{obs,0} = 5.05$, $t_{obs,1} = 2.54$, $t_{obs,2} = 2.67$, $t_{obs,3} = 0.90$, $t_{obs,4} = 1.11$; b) $\hat{y}(x_p) = 12.917$.

8. a) $\hat{y} = 14.186 - 2.041x_1 - 0.53x_2$; b) $F_{obs} = 6.918$; c) $\beta_1 \in (-2.968;\ -1.114)$, $\beta_2 \in (-1.654;\ 0.594)$; d) $R^2 = 0.634$.

9. a) $\widehat{D} = -125.56 - 4.71v + 0.046v^2$; b) $R^2 = 0.948$;

   c)

Fuente	g. l.	*SC*	*MC*	*F*
Regresión	2	43396.4	21695.2	45.37
Error	5	2391.1	478.2	
Total	7	45787.5		

10. a) Modelo cuadrático; b) $\widehat{T} = 24.14 - 2.70M + 0.21M^2$; c) $F_{obs} = 9.93$; d) $t_{0,obs} = 18.63$, $t_{1,obs} = -4.46$, $t_{2,obs} = 4.32$.

11. a) Tiempo = 10.55 – 2.02 Cant. +0.199(Cant.)^2; b) $\widehat{T} = 5.87$; c) $R^2 = 0.853$, $R_a^2 = 0.794$.

12. a) Concentr = 103.79 – 8.29 Temp. + 0.50(Temp.)^2;

b)

Fuente de variación	g. l.	SC	MC	F
Regresión	2	2571.25	1285.62	40.49
Residual	5	158.75	31.75	
Total	7	2730.00		

c) Si, modelo final: Concentr = 41.334 + 0.236(Temp.)^2.

13. Sexo: 1 si es hombre; 0 si es mujer. a) Sueldo = 7.997 + 1.147 Experiencia +1.748 Sexo; b) $F_{obs} = 9.8$; c) $t_{0,obs} = 4.48$, $t_{1,obs} = 3.64$, $t_{2,obs} = 1.61$; d) Puede decirse que no existe discriminación.

14. Sexo: 0 si es hombre; 1 si es mujer. a) $\widehat{C} = 4.63 + 0.911$ Sexo–0.223 Año; b) $R^2 = 0.817$

Fuente	g. l.	SC	MC	F
Regresión	2	1.826	0.913	11.15
Error	5	0.409	0.082	
Total	7	2.235		

c) $t_{0,obs} = 17.90$, $t_{1,obs} = 4.41$, $t_{2,obs} = -2.58$.

15. Veneno: A=0, B=1. a) Sup= 5.226 + 0.556 Edad–1.971 Ven; b) $F_{obs} = 6.55$, $R^2 = 0.652$; c) $t_{0,obs} = 6.33$, $t_{1,obs} = 1.88$, $t_{2,obs} = -3.26$; d) Sup= 6.56 – 1.86 Ven.

Capítulo 12

Sección 12.5

1. b) $\widehat{Y}_{17} = 31$, **ECM** = 186.9; c) $\widehat{Y}_{17} = 30.7$, **ECM** = 195.5; d) $\widehat{Y}_{17} = 29.7$, **ECM** = 279; e) $\widehat{Y}_{17} = 33.7$, **ECM** = 299; f) $\widehat{Y}_{17} = 34.5$, **ECM** = 280.2; g) $\widehat{Y}_{17} = 18.7$, $\widehat{Y}_{18} = 42.8$, $\widehat{Y}_{19} = 55$, $\widehat{Y}_{20} = 23.9$, **ECM** = 126.3.

2. b) $\widehat{Y}_{25} = 57.5$, **ECM** = 172.5; c) $\widehat{Y}_{25} = 56.8$, **ECM** = 153.7; d) $\widehat{Y}_{25} = 47$, **ECM** = 253.1; e) $\widehat{Y}_{25} = 59.7$, **ECM** = 243.7; f) $\widehat{Y}_{25} = 57.4$, **ECM** = 224; g) $\widehat{Y}_{25} = 75.1$, $\widehat{Y}_{26} = 54.6$, $\widehat{Y}_{27} = 50.0$, $\widehat{Y}_{28} = 71.9$, **ECM** = 124.3.

3. b) $\widehat{Y}_{16} = 117$, **ECM** = 132.7; c) $\widehat{Y}_{16} = 127$, **ECM** = 63.7; d) $\widehat{Y}_{16} = 116.2$, **ECM** = 143.7; e) $\widehat{Y}_{16} = 118..7$, **ECM** = 39.8; f) $\widehat{Y}_{16} = 126.6$, **ECM** = 83.8; g) $\widehat{Y}_{16} = 125.0$, $\widehat{Y}_{17} = 127.5$, $\widehat{Y}_{18} = 133.4$, **ECM** = 157.5.

Capítulo 13

Sección 13.3

1. a) $\mu \in (81.23; 87.17)$, $\tau \in (32\,493; 34\,867)$; b) 121.

2. a) $(738\,153; 777\,047)$; b) 192
3. a) $(326\,523; 361\,157)$; b) 24.
4. a) $(83.48; 85.32)$; b) 101.
5. a) $(4.34; 4.66)$; b) 89; c) $(6301; 6920)$.
6. a) 73; b) 196; c) $(47.48; 48.52)$.
7. a) $\mu \in (115.44; 134.56)$, $\tau \in (32\,323; 37\,677)$; b) 110.
8. a) $(0.251; 0.410)$; b) 153; c) 164.
9. a) $(0.238; 0.370)$; b) (i.) 400, (ii.) 461.
10. a) $(0.24; 0.36)$; b) 7684; c) 8160.
11. a) $(0.272; 0.408)$; b) 1228; c) 1179.
12. a) 113; b) 50; c) 351; d) 351.

Sección 13.5

1. a) $\tau \in (2\,558\,604; 2\,587\,396)$; b) $\mu \in (426.43; 431.23)$.
2. a) $(69.44; 72.11)$; b) $(8\,126\,172; 8\,438\,556)$; c) $n_1 = n_2 = n_3 = 31$; d) $n_1 = 8$, $n_2 = 23$, $n_3 = 45$; e) $n_1 = 10$, $n_2 = 24$, $n_3 = 41$.
3. a) $\mu \in (102.61; 104.48)$, $\tau \in (492\,512; 501\,488)$; b) $n_1 = n_2 = 65$; c) $n_1 = 70$, $n_2 = 65$; d) $n_1 = 33$, $n_2 = 79$.
4. a) $(608\,289; 617\,601)$; b) $n_1 = n_2 = n_3 = 79$; c) $n_1 = 93$, $n_2 = 42$, $n_3 = 86$; d) $n_1 = 70$, $n_2 = 41$, $n_3 = 101$.
5. a) $(1632.79; 1868.75)$; b) $(1\,061\,313; 1\,214\,687)$; c) $n_1 = n_2 = n_3 = 25$; d) $n_1 = 9$, $n_2 = 13$, $n_3 = 34$.
6. a) $\mu \in (725.29; 763.38)$, $\tau \in (768\,812; 809\,188)$; b) No; c) $n_1 = n_2 = n_3 = 149$; d) $n_1 = 182$, $n_2 = 154$, $n_3 = 93$; e) $n_1 = 190$, $n_2 = 121$, $n_3 = 107$.
7. $(0.349; 0.455)$.
8. a) $(0.4209; 0.5173)$; b) no; c) $n_1 = n_2 = 145$; d) $n_1 = 148$, $n_2 = 142$; e) $n_1 = 147$, $n_2 = 143$.
9. $(30.98\,\%; 39.08\,\%)$.
10. a) $(0.5705; 0.6823)$; b) $n_1 = n_2 = n_3 = 378$; c) $n_1 = 157$, $n_2 = 469$, $n_3 = 352$; d) $n_1 = 173$, $n_2 = 502$, $n_3 = 302$.
11. a) $(21.48\,\%; 31.19\,\%)$; b) 6292; c) 5972; d) 5976.
12. a) $p \in (0.5447; 0.6969)$; b) $\mu \in (5972; 6202)$; c) $n_1 = n_2 = n_3 = 58$; d) $n_1 = 84$, $n_2 = 35$, $n_3 = 56$; e) $n_1 = 86$, $n_2 = 38$, $n_3 = 51$.

Sección 13.7

1. a) $\tau \in (475\,646; 484\,354)$; b) $\mu \in (3.171; 3.229)$.
2. a) $\tau \in (889\,546; 1\,119\,228)$, $\mu \in (277.98; 349.76)$; b) 28.
3. a) (7.31;11.64); b) 98 clases.
4. $p \in (0.1; 0.3)$.
5. a) $(0.2831; 0.3733)$; b) 25.
6. a) $(0.0677; 0.2448)$; b) $(115; 416)$; c) 79.
7. a) $\mu_i \in (45.56; 85.12)$; b) $\mu_g \in (20.32; 37.12)$; c) $\mu_i \in (35.38; 106.89)$; d) $p \in (0.0272; 0.2228)$; e) $n_a = 26$, $n_b = 25$, $n_c = 29$. El tamaño óptimo es de 29 familias.

Tablas Estadísticas

1. Probabilidad acumulada de la ley de distribución normal estándar.
2. Probabilidad acumulada de $1 - \Phi(z)$ para la ley de distribución normal estándar.
3. Puntos porcentuales de la ley de distribución t de Student.
4. Puntos porcentuales de la ley de distribución χ^2.
5. Puntos porcentuales de la distribución del número de rachas.
6. Puntos porcentuales de la ley de distribución F.
   - Nivel $\alpha = 0.1$.
   - Nivel $\alpha = 0.05$.
   - Nivel $\alpha = 0.025$.
   - Nivel $\alpha = 0.01$.
7. Puntos porcentuales de la distribución de los rangos signados de Wilcoxon.
8. Puntos porcentuales de la distribución de la prueba de Kolmogorov-Smirnov.
9. Puntos porcentuales de la distribución de la prueba de Grubbs.
10. Puntos porcentuales de la distribución del coeficiente de correlación de Spearman.
11. Puntos porcentuales de la distribución de la prueba de Mann-Whitney.

Tabla 1. Probabilidad acumulada de $-\infty$ a z

para la distribución normal estándar

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Tabla 1. Probabilidad acumulada de $-\infty$ a z

para la distribución normal estándar(continuación)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998

Tabla 2. Probabilidad acumulada para $1 - \Phi(z)$

para la distribución normal estándar.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002

Tabla 3. Puntos porcentuales de la distribución t−Student

	Nivel de probabilidad (α)									
g.l.	**0.4**	**0.3**	**0.2**	**0.15**	**0.1**	**0.05**	**0.025**	**0.01**	**0.005**	**0.0025**
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706	31.821	63.656	127.321
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303	6.965	9.925	14.089
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306	2.896	3.355	3.833
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228	2.764	3.169	3.581
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201	2.718	3.106	3.497
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.428
13	0.259	0.538	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.372
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.326
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.286
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.252
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.222
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.197
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.174
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.153
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.135
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.119
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.104
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.091
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.078
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.067
27	0.256	0.531	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.057
28	0.256	0.530	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.047
29	0.256	0.530	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.038
30	0.256	0.530	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.030
35	0.255	0.529	0.852	1.052	1.306	1.690	2.030	2.438	2.724	2.996
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021	2.423	2.704	2.971
45	0.255	0.528	0.850	1.049	1.301	1.679	2.014	2.412	2.690	2.952
50	0.255	0.528	0.849	1.047	1.299	1.676	2.009	2.403	2.678	2.937
60	0.254	0.527	0.848	1.045	1.296	1.671	2.000	2.390	2.660	2.915
70	0.254	0.527	0.847	1.044	1.294	1.667	1.994	2.381	2.648	2.899
80	0.254	0.526	0.846	1.043	1.292	1.664	1.990	2.374	2.639	2.887
90	0.254	0.526	0.846	1.042	1.291	1.662	1.987	2.368	2.632	2.878
100	0.254	0.526	0.845	1.042	1.290	1.660	1.984	2.364	2.626	2.871
∞	0.675	0.525	0.675	1.037	1.282	1.645	1.960	2.327	2.576	2.808

Tabla 4. Puntos porcentuales de la distribución ji-cuadrado

g.l.	Nivel de probabilidad (α) 0.995	0.99	0.975	0.95	0.925	0.9	0.1	0.075	0.05	0.025	0.01	0.005
1	0.00	0.00	0.00	0.00	0.01	0.02	2.71	3.17	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.16	0.21	4.61	5.18	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.47	0.58	6.25	6.90	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	0.90	1.06	7.78	8.50	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.39	1.61	9.24	10.01	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	1.94	2.20	10.64	11.47	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.53	2.83	12.02	12.88	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.14	3.49	13.36	14.27	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	3.78	4.17	14.68	15.63	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.45	4.87	15.99	16.97	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.12	5.58	17.28	18.29	19.68	21.92	24.73	26.76
12	3.07	3.57	4.40	5.23	5.82	6.30	18.55	19.60	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	6.52	7.04	19.81	20.90	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.24	7.79	21.06	22.18	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	7.97	8.55	22.31	23.45	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	8.71	9.31	23.54	24.72	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	9.45	10.09	24.77	25.97	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.21	10.86	25.99	27.22	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	10.97	11.65	27.20	28.46	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	11.73	12.44	28.41	29.69	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	12.50	13.24	29.62	30.92	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	13.28	14.04	30.81	32.14	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.06	14.85	32.01	33.36	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	14.85	15.66	33.20	34.57	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	15.64	16.47	34.38	35.78	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	16.44	17.29	35.56	36.98	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	17.24	18.11	36.74	38.18	40.11	43.19	46.96	49.65
28	12.46	13.56	15.31	16.93	18.05	18.94	37.92	39.38	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	18.85	19.77	39.09	40.57	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	19.66	20.60	40.26	41.76	43.77	46.98	50.89	53.67
35	17.19	18.51	20.57	22.47	23.76	24.80	46.06	47.66	49.80	53.20	57.34	60.27
40	20.71	22.16	24.43	26.51	27.93	29.05	51.81	53.50	55.76	59.34	63.69	66.77
45	24.31	25.90	28.37	30.61	32.14	33.35	57.51	59.29	61.66	65.41	69.96	73.17
50	27.99	29.71	32.36	34.76	36.40	37.69	63.17	65.03	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	45.02	46.46	74.40	76.41	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	53.75	55.33	85.53	87.68	90.53	95.02	100.43	104.21
80	51.17	53.54	57.15	60.39	62.57	64.28	96.58	98.86	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	71.46	73.29	107.57	109.97	113.15	118.14	124.12	128.30
100	67.33	70.06	74.22	77.93	80.41	82.36	118.50	121.02	124.34	129.56	135.81	140.17
120	83.85	86.92	91.57	95.70	98.46	100.62	140.23	142.96	146.57	152.21	158.95	163.65

Tabla 5. Puntos porcentuales de la distribución del número de rachas

n	Nivel	m								
		2	3	4	5	6	7	8	9	10
2	$\alpha=0.05$	0-4	0-5	0-5	0-5	0-5	0-5	2-3	2-5	2-5
	$\alpha=0.01$	0-4	0-5	0-5	0-5	0-5	0-5	0-5	0-5	0-5
3	$\alpha=0.05$	0-5	0-6	0-6	2-7	2-7	2-7	2-7	2-7	3-7
	$\alpha=0.01$	0-5	0-6	0-7	0-7	0-7	0-7	0-7	2-7	2-7
4	$\alpha=0.05$	0-5	0-6	2-7	2-8	3-8	3-8	3-9	3-9	3-9
	$\alpha=0.01$	0-5	0-7	0-8	0-9	2-9	2-9	2-9	2-9	2-9
5	$\alpha=0.05$	0-5	2-7	2-8	3-8	3-9	3-9	3-10	4-10	4-10
	$\alpha=0.01$	0-5	0-7	0-9	2-9	2-10	2-10	2-11	3-10	3-11
6	$\alpha=0.05$	0-5	2-7	3-8	3-9	3-10	4-10	4-11	4-11	5-11
	$\alpha=0.01$	0-5	0-7	2-9	2-10	2-11	3-11	3-12	3-12	3-13
7	$\alpha=0.05$	0-5	2-7	3-8	3-9	4-10	4-11	4-12	5-12	5-12
	$\alpha=0.01$	0-5	0-7	2-9	2-10	3-11	3-12	3-12	4-13	4-14
8	$\alpha=0.05$	2-5	2-7	3-9	3-10	4-11	4-12	5-12	5-13	6-13
	$\alpha=0.01$	0-5	0-7	2-9	2-11	3-12	3-12	4-13	4-14	4-14
9	$\alpha=0.05$	2-5	2-7	3-9	4-10	4-11	5-12	5-13	6-13	6-14
	$\alpha=0.01$	0-5	2-7	2-9	3-10	3-12	4-13	4-14	4-15	5-15
10	$\alpha=0.05$	2-5	3-7	3-9	4-10	5-11	5-12	6-13	6-14	6-15
	$\alpha=0.01$	0-5	2-7	2-9	3-11	3-13	4-14	4-14	5-15	5-16
11	$\alpha=0.05$	2-5	3-7	3-9	4-11	5-12	5-13	6-14	6-14	7-15
	$\alpha=0.01$	0-5	2-7	2-9	3-11	4-13	4-14	4-15	5-16	5-17
12	$\alpha=0.05$	2-5	3-7	4-9	4-11	5-12	6-13	6-14	7-15	7-16
	$\alpha=0.01$	0-5	2-7	3-9	3-11	4-13	4-14	5-15	5-16	6-17
13	$\alpha=0.05$	2-5	3-7	4-9	4-11	5-12	6-13	6-14	7-15	8-16
	$\alpha=0.01$	0-5	2-7	3-9	3-11	4-13	5-15	5-16	6-17	6-18
14	$\alpha=0.05$	2-5	3-7	4-9	5-11	5-12	6-13	7-15	7-16	8-16
	$\alpha=0.01$	0-5	2-7	3-9	3-11	4-13	5-15	5-16	6-17	6-18
15	$\alpha=0.05$	2-5	3-7	4-9	5-11	6-13	6-14	7-15	8-16	8-17
	$\alpha=0.01$	0-5	2-7	3-9	4-11	4-13	5-15	5-16	6-17	7-18

Tabla 6. Puntos porcentuales de la distribución $F(n_1, n_2)$

Nivel de probabilidad $\alpha = 0.1$

n_2	n_1: Grados de libertad para el numerador																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	20	30	40	50	100	∞
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.47	60.71	60.90	61.07	61.22	61.74	62.26	62.53	62.69	63.01	63.32
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.40	9.41	9.41	9.42	9.42	9.44	9.46	9.47	9.47	9.48	9.49
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.22	5.21	5.20	5.20	5.18	5.17	5.16	5.15	5.14	5.13
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.91	3.90	3.89	3.88	3.87	3.84	3.82	3.80	3.80	3.78	3.76
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.28	3.27	3.26	3.25	3.24	3.21	3.17	3.16	3.15	3.13	3.11
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.92	2.90	2.89	2.88	2.87	2.84	2.80	2.78	2.77	2.75	2.72
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.68	2.67	2.65	2.64	2.63	2.59	2.56	2.54	2.52	2.50	2.47
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.52	2.50	2.49	2.48	2.46	2.42	2.38	2.36	2.35	2.32	2.29
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.40	2.38	2.36	2.35	2.34	2.30	2.25	2.23	2.22	2.19	2.16
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.30	2.28	2.27	2.26	2.24	2.20	2.16	2.13	2.12	2.09	2.06
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.23	2.21	2.19	2.18	2.17	2.12	2.08	2.05	2.04	2.01	1.97
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.17	2.15	2.13	2.12	2.10	2.06	2.01	1.99	1.97	1.94	1.90
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.12	2.10	2.08	2.07	2.05	2.01	1.96	1.93	1.92	1.88	1.85
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.07	2.05	2.04	2.02	2.01	1.96	1.91	1.89	1.87	1.83	1.80
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.04	2.02	2.00	1.99	1.97	1.92	1.87	1.85	1.83	1.79	1.76
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	2.01	1.99	1.97	1.95	1.94	1.89	1.84	1.81	1.79	1.76	1.72
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.98	1.96	1.94	1.93	1.91	1.86	1.81	1.78	1.76	1.73	1.69
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.95	1.93	1.92	1.90	1.89	1.84	1.78	1.75	1.74	1.70	1.66
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.93	1.91	1.89	1.88	1.86	1.81	1.76	1.73	1.71	1.67	1.63
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.91	1.89	1.87	1.86	1.84	1.79	1.74	1.71	1.69	1.65	1.61
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.90	1.87	1.86	1.84	1.83	1.78	1.72	1.69	1.67	1.63	1.59
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.88	1.86	1.84	1.83	1.81	1.76	1.70	1.67	1.65	1.61	1.57
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.87	1.84	1.83	1.81	1.80	1.74	1.69	1.66	1.64	1.59	1.55
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.85	1.83	1.81	1.80	1.78	1.73	1.67	1.64	1.62	1.58	1.53
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.84	1.82	1.80	1.79	1.77	1.72	1.66	1.63	1.61	1.56	1.52
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.83	1.81	1.79	1.77	1.76	1.71	1.65	1.61	1.59	1.55	1.50
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.82	1.80	1.78	1.76	1.75	1.70	1.64	1.60	1.58	1.54	1.49
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.81	1.79	1.77	1.75	1.74	1.69	1.63	1.59	1.57	1.53	1.48
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.80	1.78	1.76	1.75	1.73	1.68	1.62	1.58	1.56	1.52	1.47
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.79	1.77	1.75	1.74	1.72	1.67	1.61	1.57	1.55	1.51	1.46
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.74	1.71	1.70	1.68	1.66	1.61	1.54	1.51	1.48	1.43	1.38
50	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76	1.73	1.70	1.68	1.66	1.64	1.63	1.57	1.50	1.46	1.44	1.39	1.33
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.68	1.66	1.64	1.62	1.60	1.54	1.48	1.44	1.41	1.36	1.29
70	2.78	2.38	2.16	2.03	1.93	1.86	1.80	1.76	1.72	1.69	1.66	1.64	1.62	1.60	1.59	1.53	1.46	1.42	1.39	1.34	1.27
80	2.77	2.37	2.15	2.02	1.92	1.85	1.79	1.75	1.71	1.68	1.65	1.63	1.61	1.59	1.57	1.51	1.44	1.40	1.38	1.32	1.25
90	2.76	2.36	2.15	2.01	1.91	1.84	1.78	1.74	1.70	1.67	1.64	1.62	1.60	1.58	1.56	1.50	1.43	1.39	1.36	1.30	1.23
100	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69	1.66	1.64	1.61	1.59	1.57	1.56	1.49	1.42	1.38	1.35	1.29	1.22
∞	2.71	2.30	2.08	1.95	1.85	1.77	1.72	1.67	1.63	1.60	1.57	1.55	1.52	1.51	1.49	1.42	1.34	1.30	1.26	1.19	1.00

Tabla 6. Puntos porcentuales de la distribución $F(n_1, n_2)$

Nivel de probabilidad $\alpha = 0.05$

n_2 \ n_1: Grados de libertad para el numerador	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	20	30	40	50	100	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.0	243.9	244.7	245.4	245.9	248.0	250.1	251.1	251.8	253.0	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.40	19.41	19.42	19.42	19.43	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.73	8.71	8.70	8.66	8.62	8.59	8.58	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.89	5.87	5.86	5.80	5.75	5.72	5.70	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.66	4.64	4.62	4.56	4.50	4.46	4.44	4.41	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.98	3.96	3.94	3.87	3.81	3.77	3.75	3.71	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.55	3.53	3.51	3.44	3.38	3.34	3.32	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.26	3.24	3.22	3.15	3.08	3.04	3.02	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.05	3.03	3.01	2.94	2.86	2.83	2.80	2.76	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.89	2.86	2.85	2.77	2.70	2.66	2.64	2.59	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.76	2.74	2.72	2.65	2.57	2.53	2.51	2.46	2.41
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.66	2.64	2.62	2.54	2.47	2.43	2.40	2.35	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.58	2.55	2.53	2.46	2.38	2.34	2.31	2.26	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.51	2.48	2.46	2.39	2.31	2.27	2.24	2.19	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.45	2.42	2.40	2.33	2.25	2.20	2.18	2.12	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42	2.40	2.37	2.35	2.28	2.19	2.15	2.12	2.07	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38	2.35	2.33	2.31	2.23	2.15	2.10	2.08	2.02	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.31	2.29	2.27	2.19	2.11	2.06	2.04	1.98	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31	2.28	2.26	2.23	2.16	2.07	2.03	2.00	1.94	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.25	2.22	2.20	2.12	2.04	1.99	1.97	1.91	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.22	2.20	2.18	2.10	2.01	1.96	1.94	1.88	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23	2.20	2.17	2.15	2.07	1.98	1.94	1.91	1.85	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.24	2.20	2.18	2.15	2.13	2.05	1.96	1.91	1.88	1.82	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.22	2.18	2.15	2.13	2.11	2.03	1.94	1.89	1.86	1.80	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.20	2.16	2.14	2.11	2.09	2.01	1.92	1.87	1.84	1.78	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.12	2.09	2.07	1.99	1.90	1.85	1.82	1.76	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.17	2.13	2.10	2.08	2.06	1.97	1.88	1.84	1.81	1.74	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.15	2.12	2.09	2.06	2.04	1.96	1.87	1.82	1.79	1.73	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.08	2.05	2.03	1.94	1.85	1.81	1.77	1.71	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09	2.06	2.04	2.01	1.93	1.84	1.79	1.76	1.70	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00	1.97	1.95	1.92	1.84	1.74	1.69	1.66	1.59	1.51
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.99	1.95	1.92	1.89	1.87	1.78	1.69	1.63	1.60	1.52	1.44
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.89	1.86	1.84	1.75	1.65	1.59	1.56	1.48	1.39
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.93	1.89	1.86	1.84	1.81	1.72	1.62	1.57	1.53	1.45	1.35
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.91	1.88	1.84	1.82	1.79	1.70	1.60	1.54	1.51	1.43	1.33
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94	1.90	1.86	1.83	1.80	1.78	1.69	1.59	1.53	1.49	1.41	1.30
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.89	1.85	1.82	1.79	1.77	1.68	1.57	1.52	1.48	1.39	1.28
∞	3.84	3.00	2.61	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.79	1.75	1.72	1.69	1.67	1.57	1.46	1.40	1.35	1.25	1.00

Tabla 6. Puntos porcentuales de la distribución $F(n_1, n_2)$

Nivel de probabilidad $\alpha = 0.025$

n_2	n_1: Grados de libertad para el numerador 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	20	30	40	50	100	∞
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	973.0	976.7	979.8	982.5	984.9	993.1	1001	1006	1008	1013	1018
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.41	39.42	39.43	39.43	39.45	39.46	39.47	39.48	39.49	39.50
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.37	14.34	14.30	14.28	14.25	14.17	14.08	14.04	14.01	13.96	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.79	8.75	8.71	8.68	8.66	8.56	8.46	8.41	8.38	8.32	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.57	6.52	6.49	6.46	6.43	6.33	6.23	6.18	6.14	6.08	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.41	5.37	5.33	5.30	5.27	5.17	5.07	5.01	4.98	4.92	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.71	4.67	4.63	4.60	4.57	4.47	4.36	4.31	4.28	4.21	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.24	4.20	4.16	4.13	4.10	4.00	3.89	3.84	3.81	3.74	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.91	3.87	3.83	3.80	3.77	3.67	3.56	3.51	3.47	3.40	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.66	3.62	3.58	3.55	3.52	3.42	3.31	3.26	3.22	3.15	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.47	3.43	3.39	3.36	3.33	3.23	3.12	3.06	3.03	2.96	2.88
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.32	3.28	3.24	3.21	3.18	3.07	2.96	2.91	2.87	2.80	2.73
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.20	3.15	3.12	3.08	3.05	2.95	2.84	2.78	2.74	2.67	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.09	3.05	3.01	2.98	2.95	2.84	2.73	2.67	2.64	2.56	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	3.01	2.96	2.92	2.89	2.86	2.76	2.64	2.59	2.55	2.47	2.40
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.93	2.89	2.85	2.82	2.79	2.68	2.57	2.51	2.47	2.40	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.87	2.82	2.79	2.75	2.72	2.62	2.50	2.44	2.41	2.33	2.25
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.81	2.77	2.73	2.70	2.67	2.56	2.44	2.38	2.35	2.27	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.76	2.72	2.68	2.65	2.62	2.51	2.39	2.33	2.30	2.22	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.72	2.68	2.64	2.60	2.57	2.46	2.35	2.29	2.25	2.17	2.09
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.68	2.64	2.60	2.56	2.53	2.42	2.31	2.25	2.21	2.13	2.04
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.65	2.60	2.56	2.53	2.50	2.39	2.27	2.21	2.17	2.09	2.00
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.62	2.57	2.53	2.50	2.47	2.36	2.24	2.18	2.14	2.06	1.97
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.59	2.54	2.50	2.47	2.44	2.33	2.21	2.15	2.11	2.02	1.94
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.56	2.51	2.48	2.44	2.41	2.30	2.18	2.12	2.08	2.00	1.91
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.54	2.49	2.45	2.42	2.39	2.28	2.16	2.09	2.05	1.97	1.88
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.51	2.47	2.43	2.39	2.36	2.25	2.13	2.07	2.03	1.94	1.85
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.49	2.45	2.41	2.37	2.34	2.23	2.11	2.05	2.01	1.92	1.83
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.48	2.43	2.39	2.36	2.32	2.21	2.09	2.03	1.99	1.90	1.81
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.46	2.41	2.37	2.34	2.31	2.20	2.07	2.01	1.97	1.88	1.79
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.33	2.29	2.25	2.21	2.18	2.07	1.94	1.88	1.83	1.74	1.64
50	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32	2.26	2.22	2.18	2.14	2.11	1.99	1.87	1.80	1.75	1.66	1.55
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.22	2.17	2.13	2.09	2.06	1.94	1.82	1.74	1.70	1.60	1.48
70	5.25	3.89	3.31	2.97	2.75	2.59	2.47	2.38	2.30	2.24	2.18	2.14	2.10	2.06	2.03	1.91	1.78	1.71	1.66	1.56	1.44
80	5.22	3.86	3.28	2.95	2.73	2.57	2.45	2.35	2.28	2.21	2.16	2.11	2.07	2.03	2.00	1.88	1.75	1.68	1.63	1.53	1.40
90	5.20	3.84	3.26	2.93	2.71	2.55	2.43	2.34	2.26	2.19	2.14	2.09	2.05	2.02	1.98	1.86	1.73	1.66	1.61	1.50	1.37
100	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24	2.18	2.12	2.08	2.04	2.00	1.97	1.85	1.71	1.64	1.59	1.48	1.35
∞	5.03	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.99	1.95	1.90	1.87	1.83	1.71	1.57	1.49	1.43	1.30	1.00

Tabla 6. Puntos porcentuales de la distribución $F(n_1, n_2)$

Nivel de probabilidad $\alpha = 0.01$

n_2	n_1: Grados de libertad para el numerador																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	20	30	40	50	100	∞
1	4052	4999	5403	5625	5764	5859	5928	5981	6022	6056	6083	6106	6126	6143	6157	6209	6261	6287	6303	6334	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.41	99.42	99.42	99.43	99.43	99.45	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.13	27.05	26.98	26.92	26.87	26.69	26.50	26.41	26.35	26.24	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.45	14.37	14.31	14.25	14.20	14.02	13.84	13.75	13.69	13.58	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.96	9.89	9.82	9.77	9.72	9.55	9.38	9.29	9.24	9.13	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72	7.66	7.60	7.56	7.40	7.23	7.14	7.09	6.99	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.54	6.47	6.41	6.36	6.31	6.16	5.99	5.91	5.86	5.75	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.73	5.67	5.61	5.56	5.52	5.36	5.20	5.12	5.07	4.96	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.18	5.11	5.05	5.01	4.96	4.81	4.65	4.57	4.52	4.41	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.77	4.71	4.65	4.60	4.56	4.41	4.25	4.17	4.12	4.01	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.46	4.40	4.34	4.29	4.25	4.10	3.94	3.86	3.81	3.71	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.22	4.16	4.10	4.05	4.01	3.86	3.70	3.62	3.57	3.47	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96	3.91	3.86	3.82	3.66	3.51	3.43	3.38	3.27	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80	3.75	3.70	3.66	3.51	3.35	3.27	3.22	3.11	3.01
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67	3.61	3.56	3.52	3.37	3.21	3.13	3.08	2.98	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.62	3.55	3.50	3.45	3.41	3.26	3.10	3.02	2.97	2.86	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.46	3.40	3.35	3.31	3.16	3.00	2.92	2.87	2.76	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.43	3.37	3.32	3.27	3.23	3.08	2.92	2.84	2.78	2.68	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.30	3.24	3.19	3.15	3.00	2.84	2.76	2.71	2.60	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.29	3.23	3.18	3.13	3.09	2.94	2.78	2.69	2.64	2.54	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.24	3.17	3.12	3.07	3.03	2.88	2.72	2.64	2.58	2.48	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12	3.07	3.02	2.98	2.83	2.67	2.58	2.53	2.42	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.14	3.07	3.02	2.97	2.93	2.78	2.62	2.54	2.48	2.37	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.09	3.03	2.98	2.93	2.89	2.74	2.58	2.49	2.44	2.33	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	3.06	2.99	2.94	2.89	2.85	2.70	2.54	2.45	2.40	2.29	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	3.02	2.96	2.90	2.86	2.81	2.66	2.50	2.42	2.36	2.25	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.99	2.93	2.87	2.82	2.78	2.63	2.47	2.38	2.33	2.22	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.96	2.90	2.84	2.79	2.75	2.60	2.44	2.35	2.30	2.19	2.07
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.93	2.87	2.81	2.77	2.73	2.57	2.41	2.33	2.27	2.16	2.04
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.91	2.84	2.79	2.74	2.70	2.55	2.39	2.30	2.25	2.13	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.73	2.66	2.61	2.56	2.52	2.37	2.20	2.11	2.06	1.94	1.81
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.63	2.56	2.51	2.46	2.42	2.27	2.10	2.01	1.95	1.82	1.68
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.56	2.50	2.44	2.39	2.35	2.20	2.03	1.94	1.88	1.75	1.60
70	7.01	4.92	4.07	3.60	3.29	3.07	2.91	2.78	2.67	2.59	2.51	2.45	2.40	2.35	2.31	2.15	1.98	1.89	1.83	1.70	1.54
80	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64	2.55	2.48	2.42	2.36	2.31	2.27	2.12	1.94	1.85	1.79	1.65	1.50
90	6.93	4.85	4.01	3.53	3.23	3.01	2.84	2.72	2.61	2.52	2.45	2.39	2.33	2.29	2.24	2.09	1.92	1.82	1.76	1.62	1.46

Tabla 7. Puntos porcentuales de la distribución de los rangos signados de Wilcoxon

	Nivel de probabilidad (α)				
n	0.005	0.01	0.025	0.05	0.1
4	0	0	0	0	1
5	0	0	0	1	3
6	0	0	1	3	4
7	0	1	3	4	6
8	1	2	4	6	9
9	2	4	6	9	11
10	4	6	9	11	15
11	6	8	11	14	18
12	8	10	14	18	22
13	10	13	18	22	27
14	13	16	22	26	32
15	16	20	26	31	37
16	20	24	30	36	43
17	24	28	35	42	49
18	28	33	41	48	56
19	33	38	47	54	63
20	38	44	53	61	70
21	44	50	59	68	78
22	49	56	67	76	87
23	55	63	74	84	95
24	62	70	82	92	105
25	69	77	90	101	114
26	76	82	99	111	125
27	84	94	108	120	135
28	92	102	117	131	146
29	101	111	127	141	158
30	110	121	138	152	170
31	119	131	148	164	182
32	129	141	160	176	195
33	139	152	171	188	208
34	149	163	183	201	222
35	160	175	196	214	236
36	172	187	209	228	251
37	184	199	222	242	266
38	196	212	236	257	282
39	208	225	250	272	298
40	221	239	265	287	314
41	235	253	280	303	331
42	248	267	295	320	349
43	263	282	311	337	366
44	277	297	328	354	385
45	292	313	344	372	403
46	308	329	362	390	423
47	324	346	379	408	442
48	340	363	397	428	463
49	357	381	416	447	483
50	374	398	435	467	504

Tabla 8. Puntos porcentuales de la distribución de la prueba de Kolmogorov-Smirnov

	Nivel de probabilidad (α)						
n	**0.1**	**0.05**	**0.02**	**0.01**	**0.005**	**0.002**	**0.001**
5	0.509	0.563	0.627	0.669	0.705	0.750	0.781
6	0.468	0.519	0.577	0.617	0.653	0.696	0.725
7	0.436	0.483	0.538	0.576	0.610	0.651	0.679
8	0.410	0.454	0.507	0.542	0.574	0.614	0.641
9	0.387	0.430	0.480	0.513	0.544	0.582	0.608
10	0.369	0.409	0.456	0.489	0.519	0.555	0.580
11	0.352	0.391	0.437	0.468	0.495	0.531	0.556
12	0.338	0.375	0.419	0.449	0.477	0.510	0.534
13	0.325	0.361	0.404	0.432	0.459	0.492	0.515
14	0.314	0.349	0.390	0.418	0.444	0.475	0.498
15	0.304	0.338	0.377	0.404	0.429	0.456	0.482
16	0.295	0.327	0.366	0.392	0.416	0.446	0.468
17	0.286	0.318	0.355	0.381	0.405	0.434	0.455
18	0.279	0.309	0.346	0.371	0.394	0.422	0.442
19	0.271	0.301	0.337	0.361	0.384	0.412	0.431
20	0.265	0.294	0.329	0.352	0.375	0.402	0.421
21	0.259	0.287	0.321	0.344	0.366	0.392	0.411
22	0.253	0.281	0.314	0.337	0.358	0.384	0.402
23	0.247	0.275	0.307	0.330	0.350	0.376	0.394
24	0.242	0.269	0.301	0.323	0.343	0.368	0.386
25	0.238	0.264	0.295	0.317	0.337	0.361	0.377
26	0.233	0.259	0.290	0.310	0.330	0.354	0.371
27	0.229	0.254	0.284	0.305	0.324	0.348	0.365
28	0.225	0.250	0.279	0.300	0.319	0.342	0.358
29	0.221	0.246	0.275	0.295	0.313	0.336	0.352
30	0.218	0.242	0.270	0.290	0.308	0.331	0.347
31	0.214	0.238	0.266	0.285	0.303	0.326	0.341
32	0.211	0.234	0.262	0.281	0.299	0.321	0.336
33	0.208	0.231	0.258	0.276	0.294	0.316	0.331
34	0.215	0.227	0.254	0.273	0.290	0.311	0.326
35	0.202	0.224	0.251	0.269	0.286	0.306	0.322
36	0.199	0.221	0.247	0.265	0.282	0.303	0.318
37	0.196	0.218	0.244	0.262	0.278	0.299	0.313
38	0.194	0.215	0.241	0.258	0.275	0.295	0.309
39	0.191	0.213	0.238	0.255	0.271	0.291	0.305
40	0.189	0.210	0.235	0.252	0.268	0.288	0.302
41	0.187	0.208	0.232	0.249	0.265	0.284	0.298
42	0.185	0.205	0.229	0.246	0.262	0.281	0.295
43	0.183	0.203	0.227	0.243	0.259	0.278	0.291
44	0.181	0.201	0.224	0.241	0.256	0.275	0.288
45	0.179	0.198	0.222	0.238	0.253	0.272	0.285
46	0.177	0.196	0.219	0.235	0.250	0.269	0.282
47	0.175	0.194	0.217	0.233	0.248	0.266	0.279
48	0.173	0.192	0.215	0.231	0.245	0.263	0.276
49	0.171	0.190	0.213	0.228	0.243	0.261	0.273
50	0.170	0.188	0.211	0.226	0.240	0.258	0.271
>50	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.52}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$	$\frac{1.73}{\sqrt{n}}$	$\frac{1.85}{\sqrt{n}}$	$\frac{1.95}{\sqrt{n}}$

Tabla 9. Puntos porcentuales para la prueba de Grubbs

n	Nivel de probabilidad (α)		
	0.10	**0.05**	**0.02**
3	1.15	1.15	1.15
4	1.46	1.48	1.49
5	1.67	1.71	1.75
6	1.82	1.89	1.94
7	1.94	2.02	2.10
8	2.03	2.13	2.22
9	2.11	2.21	2.32
10	2.18	2.29	2.41
11	2.23	2.36	2.48
12	2.29	2.41	2.55
13	2.33	2.46	2.61
14	2.37	2.51	2.66
15	2.41	2.55	2.71
16	2.44	2.59	2.75
17	2.47	2.62	2.79
18	2.50	2.65	2.82
19	2.53	2.68	2.85
20	2.56	2.71	2.88
21	2.58	2.73	2.91
22	2.60	2.76	2.94
23	2.62	2.78	2.96
24	2.64	2.80	2.99
25	2.66	2.82	3.01
30	2.75	2.91	3.03

Tabla 10. Puntos porcentuales de la distribución del coeficiente de correlación de Spearman

	Nivel de probabilidad (α) - bilateral				
n	**0.005**	**0.01**	**0.02**	**0.05**	**0.1**
4					1.000
5			1.000	1.000	0.900
6	1.000	1.000	0.943	0.886	0.829
7	0.964	0.929	0.893	0.786	0.714
8	0.905	0.881	0.833	0.738	0.643
9	0.867	0.833	0.783	0.700	0.600
10	0.830	0.794	0.745	0.648	0.564
11	0.800	0.755	0.709	0.618	0.536
12	0.776	0.727	0.671	0.587	0.503
13	0.747	0.703	0.648	0.560	0.484
14	0.723	0.675	0.622	0.538	0.464
15	0.700	0.654	0.604	0.521	0.443
16	0.679	0.635	0.582	0.503	0.429
17	0.662	0.615	0.566	0.485	0.414
18	0.643	0.600	0.550	0.472	0.401
19	0.628	0.584	0.535	0.460	0.391
20	0.612	0.570	0.520	0.447	0.380
21	0.599	0.556	0.508	0.435	0.370
22	0.586	0.544	0.496	0.425	0.361
23	0.573	0.532	0.486	0.415	0.353
24	0.562	0.521	0.476	0.406	0.344
25	0.551	0.511	0.466	0.398	0.337
26	0.541	0.501	0.457	0.390	0.331
27	0.531	0.491	0.448	0.382	0.324
28	0.522	0.483	0.440	0.375	0.317
29	0.513	0.475	0.433	0.368	0.312
30	0.504	0.467	0.425	0.362	0.306
31	0.496	0.459	0.418	0.356	0.301
32	0.489	0.452	0.412	0.350	0.296
33	0.482	0.446	0.405	0.345	0.291
34	0.475	0.439	0.399	0.340	0.287
35	0.468	0.433	0.394	0.335	0.283
36	0.462	0.427	0.388	0.330	0.279
37	0.456	0.421	0.383	0.325	0.275
38	0.450	0.415	0.378	0.321	0.271
39	0.444	0.410	0.373	0.317	0.267
40	0.439	0.405	0.368	0.313	0.264
41	0.433	0.400	0.364	0.309	0.261
42	0.428	0.395	0.359	0.305	0.257
43	0.423	0.391	0.355	0.301	0.254
44	0.419	0.386	0.351	0.298	0.251
45	0.414	0.382	0.347	0.294	0.248
46	0.410	0.378	0.343	0.291	0.246
47	0.405	0.374	0.340	0.288	0.243
48	0.401	0.370	0.336	0.285	0.240
49	0.397	0.366	0.333	0.282	0.238
50	0.393	0.363	0.329	0.279	0.235

Tabla 11. Puntos porcentuales de la distribución del coeficiente de Mann-Whitney

Nivel de probabilidad $\alpha = 0.025$

	n_1																		
n_2	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
2	0	0	0	0	0	0	1	1	1	1	2	2	2	2	2	3	3	3	3
3	0	0	0	1	2	2	3	3	4	4	5	5	6	6	7	7	8	8	9
4	0	0	1	2	3	4	5	5	6	7	8	9	10	11	12	12	13	14	15
5	0	1	2	3	4	6	7	8	9	10	12	13	14	15	16	18	19	20	21
6	0	2	3	4	6	7	9	11	12	14	15	17	18	20	22	23	25	26	28
7	0	2	4	6	7	9	11	13	15	17	19	21	23	25	27	29	31	33	35
8	1	3	5	7	9	11	14	16	18	20	23	25	27	30	32	35	37	39	42
9	1	3	5	8	11	13	16	18	21	24	27	29	32	35	38	40	43	46	49
10	1	4	6	9	12	15	18	21	24	27	30	34	37	40	43	46	49	53	56
11	1	4	7	10	14	17	20	24	27	31	34	38	41	45	48	52	56	59	63
12	2	5	8	12	15	19	23	27	30	34	38	42	46	50	54	58	62	66	70
13	2	5	9	13	17	21	25	29	34	38	42	46	51	55	60	64	68	73	77
14	2	6	10	14	18	23	27	32	37	41	46	51	56	60	65	70	75	79	84
15	2	6	11	15	20	25	30	35	40	45	50	55	60	65	71	76	81	86	91
16	2	7	12	16	22	27	32	38	43	48	54	60	65	71	73	82	87	93	99
17	3	7	12	18	23	29	35	40	46	52	58	64	70	76	82	88	94	100	106
18	3	8	13	19	25	31	37	43	49	56	62	68	71	81	87	94	100	107	113
19	3	8	14	20	26	33	39	46	53	59	66	73	79	86	93	100	107	114	120
20	3	9	15	21	28	35	42	49	56	63	70	77	84	91	99	106	113	120	128

Nivel de probabilidad $\alpha = 0.01$

	n_1																		
n_2	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
2	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	2	2
3	0	0	0	0	0	1	1	2	2	2	3	3	3	4	4	5	5	5	6
4	0	0	0	1	2	2	3	4	4	5	6	6	7	8	8	9	10	10	11
5	0	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
6	0	0	2	3	4	5	7	8	9	10	12	13	14	16	16	19	20	21	23
7	0	1	2	4	5	7	8	10	12	13	15	17	18	20	22	24	25	27	29
8	0	1	3	5	7	8	10	12	14	16	18	21	23	25	27	29	31	33	35
9	0	2	4	6	8	10	12	15	17	19	22	24	27	29	32	34	37	39	41
10	0	2	4	7	9	12	14	17	20	23	25	28	31	34	37	39	42	45	48
11	0	2	5	8	10	13	16	19	23	26	29	32	35	38	42	45	48	51	54
12	0	3	6	9	12	15	18	22	25	29	32	36	39	43	47	50	54	57	61
13	1	3	6	10	13	17	21	24	28	32	36	40	44	48	52	56	60	64	68
14	1	3	7	11	14	18	23	27	31	35	39	44	48	52	57	64	66	72	74
15	1	4	8	12	16	20	25	29	34	38	43	48	52	57	62	67	71	76	81
16	1	4	8	13	17	22	27	32	37	42	47	52	57	62	67	72	77	83	88
17	1	5	9	14	19	24	29	34	39	45	50	56	61	67	72	78	83	89	94
18	1	5	10	15	20	25	31	37	42	48	54	60	66	71	77	83	89	95	101
19	2	5	10	16	21	27	33	39	45	51	57	64	70	76	83	89	95	102	108
20	2	6	11	17	23	29	35	41	48	54	61	68	74	81	88	94	101	108	115